

AD-A137 619

ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICES
ELECTRONICS PROGRAM VO. (U) TEXAS TECH UNIV LUBBOCK
INST FOR ELECTRONIC SCIENCE R SAEKS ET AL. DEC 82
N00014-76-C-1136

1/4

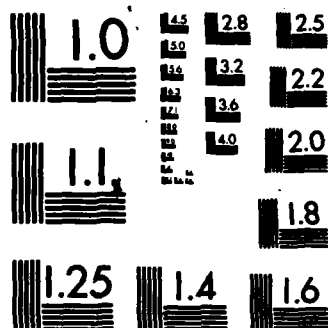
UNCLASSIFIED

F/G 9/3

NL

15

15



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

ADA137619

12

ANNUAL REVIEW OF RESEARCH

under the

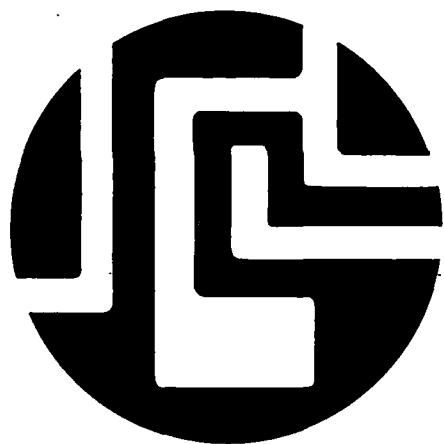
JOINT SERVICES ELECTRONICS PROGRAM

Vol 2

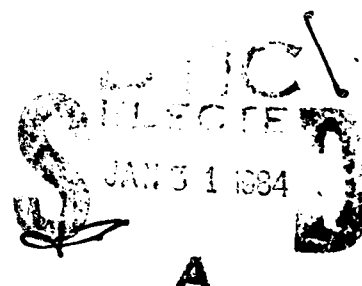
December 1982

(Publications)

N00014-76-C-1136



Institute for
Electronic Science



for information and sale only
distribution is unlimited

TEXAS TECH UNIVERSITY

Lubbock, Texas 79409

DTIC FILE COPY

→ Contents:

PUBLICATION LISTING

→ FEEDBACK SYSTEM DESIGN; R. Saeks

Feedback System Design: The Single Variate Case - Part I.....	1
Feedback System Design: The Single Variate Case - Part II....	35
Simultaneous Design of Control Systems.....	67
Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem.....	73

NONLINEAR CONTROL; L. R. Hunt

Theory of Design Using Nonlinear Transformations.....	85
Applications to Aeronautics of the Theory of Transformations of Nonlinear Systems.....	91
Design for Multi-Input Nonlinear Systems.....	101
Robustness in Nonlinear Control.....	133
Global Transformations of Nonlinear Systems.....	157
N-Dimensional Controllability with (n-1) Controls.....	165
Sufficient Conditions for Controllability.....	171
Control of Nonlinear Time-Varying Systems.....	175

NONLINEAR FAULT ANALYSIS; R. Saeks

Data base for Symbolic Network Analysis.....	183
Diagnosability of Nonlinear Circuits and Systems - Part II: Dynamical Systems.....	191

MULTIDIMENSIONAL SYSTEM THEORY; J. Murray

Feedback System Design: The Single-Variate Case - Part I.....	207
A Design Method for Two-Dimensional Recursive Digital Filters.....	241
Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem.....	249

> DETECTION AND ESTIMATION IN IMAGERY; J. Walkup

Signal Recovery from Signal Dependent Noise..... 267

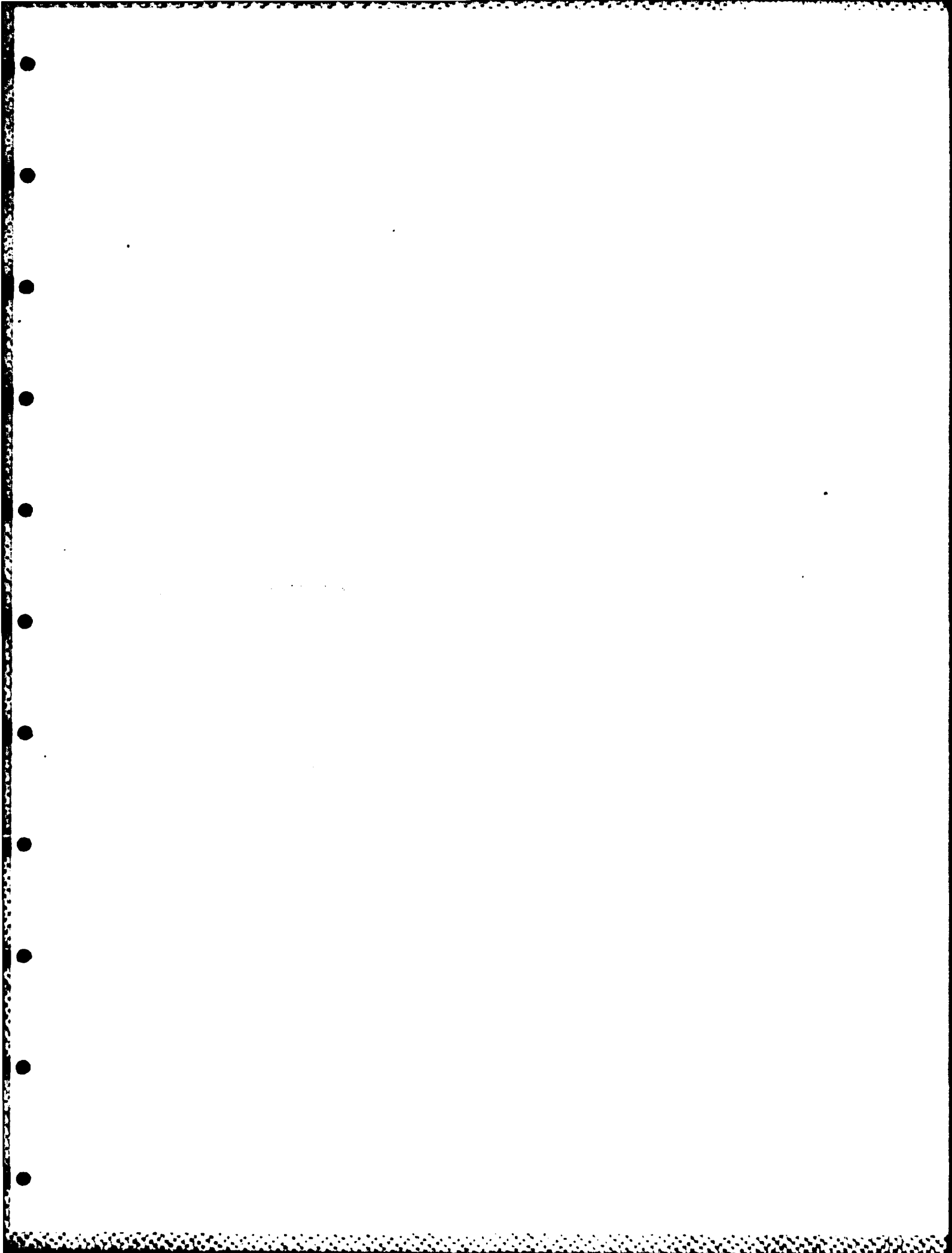
✓ POINTING AND TRACKING. T.G. Newman

The Geometry of Second Order Linear Partial Differential
Operators..... 275

FEEDBACK SYSTEM DESIGN



Accession For	
1118 GRAZI	<input checked="" type="checkbox"/>
1118 T4b	<input type="checkbox"/>
1118 1118	<input type="checkbox"/>
1118 1118	<input type="checkbox"/>
Accession/	
Availability Codes	
1118 1118	
1118 Special	
AI	



FEEDBACK SYSTEM DESIGN:

The Single-Variate Case — Part I*

R. Saeks¹, J. Murray¹, O. Chua², C. Karmokolias³,
and A. Iyer¹

Abstract. A recently developed algebraic approach to the feedback system design problem is reviewed via the derivation of the theory in the single-variate case. This allows the simple algebraic nature of the theory to be brought to the fore while simultaneously minimizing the complexities of the presentation. Rather than simply giving a single solution to the prescribed design problem we endeavor to give a complete parameterization of the set of compensators which meet specifications. Although this might at first seem to complicate our theory it, in fact, opens the way for a sequential approach to the design problem in which one parameterizes the subset of those compensators which meet the second specification...etc. Specific problems investigated include feedback system stabilization, the tracking and disturbance rejection problem, robust design, transfer function design, pole placement, simultaneous stabilization, and stable stabilization.

1. Introduction

In 1976 Youla, Bongiorno, and Jabr published two, now classical, papers [23,24] in which a complete parameterization of the set of stabilizing compensators for a multivariate feedback system was obtained. In the ensuing years this work, which is often termed the YBJ theory, has led to the development of an entirely new approach to the *feedback system design problem*. Indeed, their stabilization theory has been extended to include:

- (i) the tracking and disturbance rejection problem
- (ii) robust design algorithms
- (iii) design with a proper or stable compensator
- (iv) transfer function design

*Received September 16, 1981; revised November 10, 1981. This research was supported by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

¹Department of Electrical Engineering, Texas Tech University, Lubbock, Texas 79409, USA.

²Presently with Honeywell, Inc., Phoenix, Arizona.

³Presently with Kearfott Division, Singer, Inc., Little Falls, New Jersey.

- (v) pole placement
- (vi) simultaneous stabilization

Moreover, much of the work has been extended to the case of general linear systems; distributed, time-varying, multidimensional, etc.; by formulating it in an abstract ring theoretic [8,10] or algebro-geometric setting [17]. Unfortunately, these generalizations have been achieved at the cost of increasing the complexity of the theory and, as such, the simple algebraic character of the work has been obscured.

The purpose of the present paper is to survey this literature in such a way as to illustrate the simplicity of the theory. To this end the presentation is restricted to the single-variate case wherein a simple algebraic theory is possible. Indeed, by so doing we are able to give simple single-variate algebraic derivations for several results whose true character has hitherto been obscured by the abstract ring theoretic or multivariable theory.

The key to our theory is a three step design philosophy

- (i) stabilization
- (ii) achievement of design constraints
- (iii) optimization of system performance

First and foremost, a feedback system must be *stable* and, as such, the first step in the design process is the *parameterization of all stabilizing compensators* for the given plant. Although it might suffice to specify a single stabilizing compensator if our goal was simply to design a stable system, in practice stabilization is only the first step of the design process. As such, we must characterize all stabilizing compensators if we are to choose among the stabilizing compensators to find one which also achieves the design constraints and/or optimizes system performance. A complete parameterization of the set of stabilizing compensators for the given plant is thus obtained as a first step in the design process. Indeed, the parameterization is chosen in such a way that the various feedback system gains are linear (affine) in the resultant design parameter, thereby setting the stage for the choice of a design parameter which also achieves the *design constraints* and/or *optimizes some measure of system performance*.

Once the stabilizing compensators have been characterized, step two of the design process is to choose a subset of the stabilizing compensators which also achieve the prescribed design constraints; tracking and disturbance rejection, transfer function specification, robustness, etc. Finally, if any remaining design latitude exists after the design constraints have been met it may be used to optimize some measure of system performance; sensitivity, energy consumption, etc.

The paper is divided into two parts dealing with the classical *asymptotic design problems*: stabilization, tracking, and disturbance rejection; and a survey of modern frequency domain design; robust design, pole placement, simultaneous design, respectively. In the remainder of this introduction the

fractional representation theory for a single-variate system is developed. The key to this theory lies with the representation of a rational function as a ratio of stable rational functions rather than as a ratio of polynomials. Such a formulation opens the door to the desired generalization, wherein stability is well defined even though no analog of a polynomial exists. Moreover, it yields what we believe to be a more *natural concept of coprimeness* in which only cancellations between (closed) right half-plane zeros are forbidden. Indeed, the (strict) left half-plane plays only a minimal role in the theory.

In Section 2 a derivation of the *YBJ stabilization theory* is formulated in terms of a stable coprime fractional representation. Although this derivation has appeared before [4,8], even in the single-variate case, a complete proof is given because of its fundamental nature to the remainder of the work. Indeed, the proof technique introduced here is repeated, in one form or another, throughout the paper.

Sections 3-5 are devoted to the *tracking and disturbance rejection problems* [4]. Unlike the stabilization problem a solution to these problems may fail to exist. Necessary and sufficient conditions for the existence of a solution are, however, given in the form of appropriate coprimeness criteria and a complete parameterization of the required set of compensators is obtained when these criteria are satisfied.

Part II of the paper begins with Section 6 in which the problem of *robust design* is taken up. Unlike the stabilization problem for which every solution is robust a solution to the tracking and/or disturbance rejection problem may fail to be robust. Surprisingly, however, whenever these problems are solvable they are robustly solvable. As such, beginning with the same coprimeness criteria used in the non-robust case we give an explicit parameterization for the set of compensators which robustly solve the tracking and disturbance rejection problems. This result, however, only applies to our single-variate case. In the general multivariate case a robust solution may fail to exist even though a non-robust solution exists [11].

In Section 7 the problem of designing a compensator which simultaneously stabilizes a feedback system and realizes a prescribed input-output feedback system gain is investigated. The required existence criteria for this *transfer function design problem* are formulated in terms of a divisibility condition in the ring of stable transfer functions. This is followed in Section 8 by an investigation of the *pole placement problem* in which one desires to construct a compensator which will simultaneously stabilize a system and place the poles of its input-output gain at prescribed points in the left half-plane. Interestingly, the extent to which this end can be achieved is determined precisely by the "degree" to which the plant fails to be miniphase.

In Section 9 the problem of designing a compensator which simultaneously stabilizes two distinct plants is solved. Although this *two plant problem* is a very special case of the general *simultaneous design problem* it

is the one example of the problem for which a definitive frequency domain design criterion exists [7] and is thus indicative of the direction of future research in this area. Moreover, the problem of stabilizing a feedback system with a *stable compensator* [25] proves to be a special case of this two plant problem which is developed in Section 10.

Section 11 is devoted to a short discussion of the "*optimization problem*" associated with step three of our design process. Since the specific optimization one might choose to undertake is dependent on the physical system under study and its application the development in this section concentrates on the interface between our theory and the optimization process, without going into specifics.

Finally, Section 12, is devoted to an *historical overview* of the theory and a discussion of the various *generalizations and extensions* which thus far have been formulated.

Our system will be described by a rational function

$$r(s) = \frac{p(s)}{q(s)} \quad (1.1)$$

Such a system is said to be *stable* if its poles lie in the (strict) left half-plane. Since the point at infinity is taken to lie on the imaginary axis this implies that $r(s)$ is stable if and only if it is a proper rational function and $q(s)$ is a (strictly) Hurwitz polynomial.

A *fractional representation* for $r(s)$ is a factorization of $r(s)$ in the form

$$r(s) = \frac{n_r(s)}{d_r(s)} \quad (1.2)$$

where both $n_r(s)$ and $d_r(s)$ are stable and $d_r(s) \neq 0$. If one is given a *polynomial fractional representation* for $r(s)$ such as in equation 1.1 then one can take

$$n_r(s) = \frac{p(s)}{m(s)} \quad (1.3)$$

and

$$d_r(s) = \frac{q(s)}{m(s)} \quad (1.4)$$

where $m(s)$ is any Hurwitz polynomial such that the order of $m(s)$ equals the order of $r(s)$ verifying the existence of the required fractional representation [11].

We say that the fractional representation $r(s) = n_r(s)/d_r(s)$ is *coprime* if there exist stable rational functions, $u_r(s)$ and $v_r(s)$, such that

$$u_r(s)n_r(s) + v_r(s)d_r(s) = 1 \quad (1.5)$$

Recall that for polynomials 1.5 is equivalent to the requirement that $n_r(s)$ and $d_r(s)$ do not have any common zeros [2]. In our case, however, where we are dealing with stable rational functions, Equation 1.5 implies that $n_r(s)$ and $d_r(s)$ have no common (closed) right half-plane zeros and conversely [4,11]. Although this represents a departure from classical control theory only right half-plane zeros cause instability and, as such, it is appropriate that only right half-plane pole zero cancellations be forbidden.

Unlike the classical polynomial fractional representation theory wherein the units are the constant functions in our theory the units are the *miniphase* rational functions which are stable and admit a stable inverse. That is, if $r(s) = p(s)/q(s)$ then $p(s)$ and $q(s)$ are both Hurwitz polynomials of the same order. As such, the classical theorems for polynomial fractional representations may be reformulated in our setting with the units taken to be miniphase rational functions as follows.

1. Property. Let $r(s) = n_r(s)/d_r(s)$ be coprime fractional representation for $r(s)$ and assume that $n_r(s)$ and $d_r(s)$ admit a common divisor, $k(s)$, such that

$$d_r(s) = y_r(s)k(s)$$

and

$$n_r(s) = x_r(s)k(s)$$

where $y_r(s)$, $x_r(s)$ and $k(s)$ are stable. Then $k(s)$ is miniphase. That is, the only common divisor of a coprime fractional representation is a unit.

Proof. Since $d_r(s) = y_r(s)k(s)$ and $n_r(s) = x_r(s)k(s)$ are coprime there exist stable $u_r(s)$ and $v_r(s)$ such that

$$1 = u_r(s)n_r(s) + v_r(s)d_r(s) = [u_r(s)x_r(s) + v_r(s)y_r(s)]k(s) \quad (1.6)$$

showing that $[u_r(s)x_r(s) + v_r(s)y_r(s)]$ is a stable inverse for $k(s)$ and hence verifying that $k(s)$ is miniphase.

2. Property. Let $r(s) = n(s)/d(s)$ be a fractional representation for $r(s)$ and let $r(s) = x_r(s)/y_r(s)$ be a coprime fractional representation for $r(s)$. Then there exists a stable $k(s)$ such that

$$n_r(s) = x_r(s)k(s)$$

and

$$d_r(s) = y_r(s)k(s)$$

Proof. Given the two fractional representations let $k(s) = d_r(s)/y_r(s)$. Then clearly $d_r(s) = y_r(s)k(s)$ while

$$n(s) = \frac{n_r(s)}{d_r(s)} d_r(s) = r(s)d(s) = \frac{x_r(s)}{y_r(s)} d_r(s) = x_r(s)k(s) \quad (1.7)$$

showing that $k(s)$ is a common factor of $n_r(s)$ and $d_r(s)$. It thus remains to verify that $k(s)$ is stable. To this end recall that since $x_r(s)/y_r(s)$ is coprime there exists stable $u_r(s)$ and $v_r(s)$ such that

$$[u_r(s)x_r(s) + v_r(s)y_r(s)] = 1 \quad (1.8)$$

hence

$$\begin{aligned} k(s) &= \frac{d_r(s)}{y_r(s)} = [u_r(s)x_r(s) + v_r(s)y_r(s)] \frac{d_r(s)}{y_r(s)} \\ &= \frac{u_r(s)x_r(s)d_r(s)}{y_r(s)} + v_r(s)d_r(s) = u_r(s)r(s)d_r(s) + v_r(s)d_r(s) \quad (1.9) \\ &= \frac{u_r(s)n_r(s)d_r(s)}{d_r(s)} + v_r(s)d_r(s) = u_r(s)n_r(s) + v_r(s)d_r(s) \end{aligned}$$

showing that $k(s)$ is stable since we have expressed it as a sum of products of stable rational functions.

Note that since a coprime fractional representation always exists [11] for $r(s)$ Property 2 implies that any pair of stable rational functions, $x_r(s)$ and $y_r(s)$, can be expressed in the form $x_r(s) = n_r(s)k(s)$ and $y_r(s) = d_r(s)k(s)$ where $n_r(s)$ and $d_r(s)$ are coprime stable rational functions and $k(s)$ is stable. As such, $k(s)$ represents a *greatest common divisor* for $x_r(s)$ and $y_r(s)$ which is unique up to a miniphase factor via Property 1.

3. Property. Let $r(s) = n_r(s)/d_r(s)$ be a coprime fractional representation for $r(s)$. Then $r(s)$ is stable if and only if $d_r(s)$ is miniphase.

Proof. If $d_r(s)$ is miniphase then $1/d_r(s)$ is stable and hence $r(s)$, being the product of two stable functions is stable. Conversely, if $r(s)$ is stable then we may express $n_r(s)$ and $d_r(s)$ in the form

$$n_r(s) = r(s)d_r(s) \quad (1.10)$$

and

$$d_r(s) = 1 d_r(s) \quad (1.11)$$

showing that $d_r(s)$ is a common factor of the coprime rational functions $n_r(s)$ and $d_r(s)$. As such Property 1 implies that $d_r(s)$ is miniphase as was to be shown.

4. Example. Consider the rational function

$$r(s) = \left[\frac{(s+1)}{(s^2-4)} \right] = \frac{\left[\frac{(s+1)}{(s+1)^2} \right]}{\left[\frac{(s-2)}{(s+2)} \right]} = \frac{n_r(s)}{d_r(s)} \quad (1.12)$$

Now, $n_r(s)$ has zeros at $s = -1$ and $s = \infty$ while $d_r(s)$ has a zero at $s = 2$. As such, $n_r(s)$ and $d_r(s)$ have no common right half-plane zeros thus implying the existence of stable rational functions $u_r(s)$ and $v_r(s)$ such that $[u_r(s)n_r(s) + v_r(s)d_r(s)] = 1$. Indeed,

$$\begin{aligned} \left[\frac{16}{3} \right] \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right] \left[\frac{(s-2)}{(s+2)} \right] \\ = u_r(s)n_r(s) + v_r(s)d_r(s) = 1 \end{aligned} \quad (1.13)$$

Note that unlike the case of a polynomial fractional representation the existence of common (strict) left half-plane zeros does not preclude coprimeness. Indeed, an alternative coprime fractional representation for the above rational function takes the form

$$r(s) = \left[\frac{(s+1)}{(s^2-4)} \right] = \frac{\left[\frac{(s+1)^2}{(s+2)^3} \right]}{\left[\frac{(s-2)(s+1)}{(s+2)^2} \right]} = \frac{n'_r(s)}{d'_r(s)} \quad (1.14)$$

where $n'_r(s)$ and $d'_r(s)$ have a common zero at $s = -1$. There are, however, still coprime since

$$\begin{aligned} \left[\frac{16(s+2)}{3(s+1)} \right] \left[\frac{(s+1)^2}{(s+2)^3} \right] + \left[\frac{(s+2/3)}{(s+1)} \right] \left[\frac{(s-2)(s+1)}{(s+2)^2} \right] \\ = u'_r(s)n'_r(s) + v'_r(s)d'_r(s) = 1 \end{aligned} \quad (1.15)$$

2. Stabilization

The basic feedback system with which we deal is shown in Figure 1.

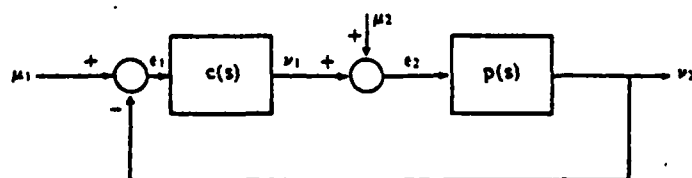


Figure 1. Basic feedback system.

For this system the usual algebraic manipulations [8] will yield the *feedback system gains*

$$\begin{bmatrix} e_1(s) \\ e_2(s) \end{bmatrix} = \begin{bmatrix} h_{e_1\mu_1}(s) & h_{e_1\mu_2}(s) \\ h_{e_2\mu_1}(s) & h_{e_2\mu_2}(s) \end{bmatrix} \begin{bmatrix} \mu_1(s) \\ \mu_2(s) \end{bmatrix} \quad (2.1)$$

where

$$\begin{bmatrix} h_{e_1\mu_1}(s) & h_{e_1\mu_2}(s) \\ h_{e_2\mu_1}(s) & h_{e_2\mu_2}(s) \end{bmatrix} = \begin{bmatrix} \frac{1}{1+p(s)c(s)} & \frac{-p(s)}{1+p(s)c(s)} \\ \frac{c(s)}{1+p(s)c(s)} & \frac{1}{1+p(s)c(s)} \end{bmatrix} \quad (2.2)$$

while

$$\begin{bmatrix} v_1(s) \\ v_2(s) \end{bmatrix} = \begin{bmatrix} h_{v_1\mu_1}(s) & h_{v_1\mu_2}(s) \\ h_{v_2\mu_1}(s) & h_{v_2\mu_2}(s) \end{bmatrix} \begin{bmatrix} \mu_1(s) \\ \mu_2(s) \end{bmatrix} \quad (2.3)$$

where

$$\begin{bmatrix} h_{v_1\mu_1}(s) & h_{v_1\mu_2}(s) \\ h_{v_2\mu_1}(s) & h_{v_2\mu_2}(s) \end{bmatrix} = \begin{bmatrix} h_{e_2\mu_1}(s) & h_{e_2\mu_2}(s) - 1 \\ 1 - h_{e_1\mu_1}(s) & -h_{e_1\mu_2}(s) \end{bmatrix} \\ = \begin{bmatrix} \frac{c(s)}{1+p(s)c(s)} & \frac{-p(s)c(s)}{1+p(s)c(s)} \\ \frac{p(s)c(s)}{1+p(s)c(s)} & \frac{p(s)}{1+p(s)c(s)} \end{bmatrix} \quad (2.4)$$

Of course, the system is said to be stable if each of the eight feedback system gains of equations 2.1 through 2.4 is stable. Since the input/output gains, $h_{\mu\mu}$, are expressed in terms of the input/plant-input gains, $h_{e\mu}$, via equation 2.4 this will be this case if and only if the input/plant-input gains are all stable.

For our stabilization theory we assume that a coprime fractional representation for the plant is given in the form

$$p(s) = \frac{n_p(s)}{d_p(s)} \quad (2.5)$$

where $n_p(s)$ and $d_p(s)$ are stable, $d_p(s)$ is not identically zero, and there exists stable $u_p(s)$ and $v_p(s)$ such that

$$u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (2.6)$$

In our single-variate setting every plants admits such a representation and hence we may assume 2.5 and 2.6 without loss of generality. Our goal is to characterize the set of compensators, represented in the form

$$c(s) = \frac{n_c(s)}{d_c(s)} \quad (2.7)$$

where $n_c(s)$ and $d_c(s)$ are stable and $d_c(s)$ is not identically zero, which stabilize the feedback system. Of course, we would also like 2.7 to be a coprime fractional representation. Indeed, so as to prevent (right half-plane)

FEEDBACK SYSTEM DESIGN

pole-zero cancellation between $p(s)$ and $c(s)$ we require the stronger condition that

$$p(s)c(s) = \frac{n_p(s)n_c(s)}{d_p(s)d_c(s)} \quad (2.8)$$

be a coprime fractional representation.

Substituting the fractional representations $p(s) = n_p(s)/d_p(s)$ and $c(s) = n_c(s)/d_c(s)$ into 2.2 and 2.4 now yields

$$\begin{bmatrix} h_{1\mu_1}(s) & h_{1\mu_2}(s) \\ h_{2\mu_1}(s) & h_{2\mu_2}(s) \end{bmatrix} = \begin{bmatrix} \frac{d_p(s)d_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} & \frac{-n_p(s)d_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} \\ \frac{d_p(s)n_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} & \frac{d_p(s)d_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} \end{bmatrix} \quad (2.9)$$

and

$$\begin{bmatrix} h_{1\mu_1}(s) & h_{1\mu_2}(s) \\ h_{2\mu_1}(s) & h_{2\mu_2}(s) \end{bmatrix} = \begin{bmatrix} \frac{d_p(s)n_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} & \frac{-n_p(s)n_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} \\ \frac{n_p(s)n_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} & \frac{n_p(s)d_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} \end{bmatrix} \quad (2.10)$$

Since the fractional representation for $p(s)c(s)$ given in 2.8 is coprime there exist stable $p(s)$ and $q(s)$ such that

$$p(s)[d_p(s)d_c(s)] + q(s)[n_p(s)n_c(s)] = 1 \quad (2.11)$$

hence

$$[p(s) - q(s)][d_p(s)d_c(s)] + q(s)[d_p(s)d_c(s) + n_p(s)n_c(s)] = 1 \quad (2.12)$$

showing that the fractional representation for $h_{1\mu_1}(s)$ given in 2.9 is coprime. As such, it follows from Property 1 that $h_{1\mu_1}(s)$ is stable if and only if the common denominator, $[d_p(s)d_c(s) + n_p(s)n_c(s)]$ is miniphase. Moreover, since $h_{1\mu_1}(s)$ must be stable for the feedback system to be stable it follows that $[d_p(s)d_c(s) + n_p(s)n_c(s)]$ must be miniphase for the system to be stable. Conversely, if this common denominator is miniphase the system is clearly stable via 2.9 and 2.10. Moreover, if the common denominator is miniphase

$$\begin{aligned} & [d_p(s)d_c(s) + n_p(s)n_c(s)]^{-1} [d_p(s)d_c(s)] \\ & + [d_p(s)d_c(s) + n_p(s)n_c(s)]^{-1} [n_p(s)n_c(s)] + 1 \end{aligned} \quad (2.13)$$

showing that the corresponding fractional representation for $p(s)c(s)$ is coprime. We have therefore proven the following.

5. Property. Let $p(s) = n_p(s)/d_p(s)$ be a coprime fractional representation for $p(s)$ and let $c(s) = n_c(s)/d_c(s)$ be a fractional representation for $c(s)$. Then the feedback system is stable and $p(s)c(s) = [n_p(s)n_c(s)]/[d_p(s)d_c(s)]$ is coprime if and only if $[d_p(s)d_c(s) + n_p(s)n_c(s)]$ is miniphase.

Consistent with Property 5 the goal of the feedback system stabilization problem is to characterize the compensators, $c(s) = n_c(s)/d_c(s)$, such that $[d_p(s)d_c(s) + n_p(s)n_c(s)]$ is miniphase given the coprime fractional representation

$$p(s) = \frac{n_p(s)}{d_p(s)} \quad (2.14)$$

where

$$u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (2.15)$$

for some stable $u_p(s)$ and $v_p(s)$.

Stabilization Theorem: For the feedback system of Figure 1 let the plant have a coprime fractional representation as per equation 2.14 and 2.15. Then for any stable $w(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero the compensator

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]} = \frac{n_c(s)}{d_c(s)}$$

stabilizes the feedback system and yields a coprime fractional representation on $p(s)c(s) = [n_p(s)n_c(s)]/[d_p(s)d_c(s)]$. Conversely, every such stabilizing compensator is of this form for some stable $w(s)$.

Proof. According to Property 5 it suffices to characterize the class of stable $n_c(s)$ and $d_c(s)$ such that

$$d_p(s)d_c(s) + n_p(s)n_c(s) = k(s) \quad (2.16)$$

where $k(s)$ is an arbitrary miniphase function. To this end we will attempt to compute all possible stable solutions to equation 2.16. Multiplying equation 2.15 through by $k(s)$ yields

$$[k(s)u_p(s)]n_p(s) + [k(s)v_p(s)]d_p(s) = k(s) \quad (2.17)$$

verifying that

$$n_c^p(s) = k(s)u_p(s) \quad (2.18)$$

and

$$d_c^p(s) = k(s)v_p(s) \quad (2.19)$$

are particular solutions to equation 2.16. On the other hand

FEEDBACK SYSTEM DESIGN

$$d_p(s) [n_p(s)r(s)] + n_p(s) [-d_p(s)r(s)] = 0 \quad (2.20)$$

for all stable $r(s)$ showing that

$$n_c^h(s) = -d_p(s)r(s) \quad (2.21)$$

and

$$d_c^h(s) = n_p(s)r(s) \quad (2.22)$$

are homogeneous solutions to 2.16 for all stable $r(s)$. It remains to show that 2.21 and 2.22 represent all homogeneous solutions. To this end let $\underline{n}_c^h(s)$ and $\underline{d}_c^h(s)$ represent arbitrary stable homogeneous solutions to 2.16. That is

$$d_p(s)\underline{d}_c^h(s) + n_p(s)\underline{n}_c^h(s) = 0 \quad (2.23)$$

in which case we will show that they take the form of 2.21 and 2.22. As a candidate for $r(s)$ let us take $r(s) = -\underline{n}_c^h(s) / d_p(s)$ in which case we have

$$\underline{n}_c^h(s) = -d_p(s)r(s) \quad (2.24)$$

verifying 2.21 and

$$\underline{d}_c^h(s) = \frac{-n_p(s)\underline{n}_c^h(s)}{d_p(s)} = n_p(s)r(s) \quad (2.25)$$

verifying 2.22. It thus remains to show that $r(s) = -\underline{n}_c^h(s) / d_p(s)$ is stable for which we have

$$\begin{aligned} r(s) &= -\frac{\underline{n}_c^h(s)}{d_p(s)} = -\frac{\underline{n}_c^h(s)}{d_p(s)} [u_p(s)n_p(s) + v_p(s)d_p(s)] \\ &= -\frac{\underline{n}_c^h(s)n_p(s)u_p(s)}{d_p(s)} - \underline{n}_c^h(s)v_p(s) = \frac{d_p(s)\underline{d}_c^h(s)u_p(s)}{d_p(s)} - \underline{n}_c^h(s)v_p(s) \\ &= \underline{d}_c^h(s)u_p(s) - \underline{n}_c^h(s)v_p(s) \end{aligned} \quad (2.26)$$

showing that $r(s)$ is stable since it is expressed as the sum of products of stable rational functions. As such, the entire solution space for equation 2.16 takes the form

$$n_c(s) = n_c^h(s) + n_c^p(s) = -r(s)d_p(s) + k(s)u_p(s) \quad (2.27)$$

and

$$d_c(s) = d_c^h(s) + d_c^p(s) = r(s)n_p(s) + k(s)v_p(s) \quad (2.28)$$

where $k(s)$ is an arbitrary miniphase function and $r(s)$ is an arbitrary stable function.

Now, assuming that $r(s)$ and $k(s)$ are chosen so that $d_c(s)$ is not identically zero we obtain the desired set of compensators in the form

$$c(s) = \frac{[-r(s)d_p(s) + k(s)u_p(s)]}{[r(s)n_p(s) + k(s)v_p(s)]}$$

$$= \frac{[-r(s)/k(s)]d_p(s) + u_p(s)}{[r(s)/k(s)]n_p(s) + v_p(s)} = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]} \quad (2.29)$$

where $w(s) = r(s)/k(s)$ spans the set of stable rational functions such that $[w(s)n_p(s) + v_p(s)]$ is not identically zero.

In addition to giving a complete parameterization of the stabilizing compensators if one views $w(s)$ rather than $c(s)$ as the underlying design parameter for our feedback system the expressions for the various feedback system gains are greatly simplified. This follows by observing that the compensator of the theorem yields the common denominator

$$[d_p(s)d_c(s) + n_p(s)n_c(s)] = 1 \quad (2.30)$$

(since we have divided $k(s)$ out of the expression for $c(s)$). As such, the denominators in Equations 2.9 and 2.10 drop out yielding the following expressions for the feedback system gains which are linear (actually affine) in the design parameter $w(s)$.

6. Corollary. The feedback system gains which result from the use of the compensator of the stabilization theorem take the form

$$\begin{bmatrix} h_{1p1}(s) & h_{1p2}(s) \\ h_{2p1}(s) & h_{2p2}(s) \end{bmatrix} = \begin{bmatrix} w(s)n_p(s)d_p(s) + v_p(s)d_p(s) & -w(s)n_p^2(s) - v_p(s)n_p(s) \\ -w(s)d_p^2(s) + u_p(s)d_p(s) & w(s)n_p(s)d_p(s) + v_p(s)d_p(s) \end{bmatrix}$$

and

$$\begin{bmatrix} h_{1p1}(s) & h_{1p2}(s) \\ h_{2p1}(s) & h_{2p2}(s) \end{bmatrix} = \begin{bmatrix} -w(s)d_p^2(s) + u_p(s)d_p(s) & w(s)n_p(s)d_p(s) - v_p(s)n_p(s) \\ -w(s)n_p(s)d_p(s) + u_p(s)n_p(s) & w(s)n_p^2(s) + v_p(s)n_p(s) \end{bmatrix}$$

Proof. These relationships result immediately upon substituting 2.30 and the expressions for $n_c(s)$ and $d_c(s)$ of the theorem into equations 2.9 and 2.10.

The theorem gives a complete parameterization of the stabilizing compensators for our feedback system modulo the requirement that $w(s)n_p(s) + v_p(s)$ not be identically zero. Needless to say, in our single-variate case this requirement is trivially verified. Moreover, $w(s)n_p(s) + v_p(s)$ is always non-zero for some $w(s)$ hence the existence of

a stabilizing compensator for any single-variate plant is guaranteed.

7. Corollary. Every single-variate plant admits a stabilizing compensator.

Proof: Consistent with the theorem it suffices to verify the existence of a stable $w(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. Indeed, either $w(s) = -1$ or $w(s) = 0$ suffices. If $w(s) = -1$ fails, i.e.,

$$d_c(s) = -n_p(s) + v_p(s) = 0 \quad (2.31)$$

then the coprimeness equality

$$1 = u_p(s)n_p(s) + v_p(s)d_p(s) = [u_p(s) + d_p(s)]v_p(s) \quad (2.32)$$

implies that $v_p(s)$ is miniphase, since $[u_p(s) + d_p(s)]$ is a stable inverse for $v_p(s)$, in which case

$$d_c(s) = [0]n_p(s) + v_p(s) = v_p(s) \quad (2.33)$$

is not identically zero.

Note that the above result is contingent on the existence of a coprime fractional representation for the plant and therefore may fail in the various generalized settings to which the theory can be extended [8]. It does, however, hold in the multivariate case wherein a coprime fractional representation is also assured to exist [11].

Occasionally, one desires to design a compensator which is a proper rational function; i.e., $c(\infty) < \infty$; rather than simply asking for a stabilizing compensator [2, 11]. Now, $c(s) = n_c(s)/d_c(s)$ is coprime via Equation 2.30 hence $n_c(\infty)$ and $d_c(\infty)$ are not both simultaneously zero. On the other hand $n_c(\infty) < \infty$ since $n_c(s)$ is stable hence

$$c(\infty) = \frac{n_c(\infty)}{d_c(\infty)} < \infty \quad (2.34)$$

if and only if $d_c(\infty) = w(\infty)n_p(\infty) + v_p(\infty) \neq 0$. Of course, in this case $w(s)n_p(s) + v_p(s)$ is not identically zero showing that the proper stabilizing compensators take the form

$$c(s) = \frac{-w(s)d_p(s) + u_p(s)}{w(s)n_p(s) + v_p(s)} \quad (2.35)$$

where $w(s)$ is stable and

$$w(\infty)n_p(\infty) + v_p(\infty) \neq 0 \quad (2.36)$$

We may now consider two cases. First, if the plant is strictly proper, $p(\infty) = 0$, then $n_p(\infty) = 0$ and since $n_p(s)$ and $v_p(s)$ are coprime via 2.15 implies that $v_p(\infty) \neq 0$. As such, 2.36 is satisfied for all stable $w(s)$. On the other hand, if the plant is not strictly proper, $p(\infty) \neq 0$, then $n_p(\infty) \neq 0$ in which case 2.36 reduces to $w(\infty) \neq -v_p(\infty)/n_p(\infty)$. We have thus verified the following corollaries [10, 11].

8. Corollary. If $p(s)$ is strictly proper then the set of compensators given by the theorem are all proper and well defined for every stable $w(s)$.

9. Corollary. If $p(s)$ is not strictly proper then the compensators given by the theorem are well defined and proper if and only if

$$w(\infty) \neq \frac{-v_p(\infty)}{n_p(\infty)}$$

Finally, rather than simply looking for a proper compensator we may desire to design a stable compensator [25]. Although such a compensator does not, in general, exist, a criterion for the existence of a stable stabilizing compensator and an algorithm for its construction is given in Section 10 as a corollary to the simultaneous stabilization theorem. The result is, however, far from elementary and no parameterization of the space of such compensators is known [17].

10. Example. For the plant of Example 4 with the coprime fractional representation of of Equations 1.12 and 1.13 the required set of stabilizing compensators take the form

$$c(s) = \frac{-w(s) \left[\frac{(s-2)}{(s+2)} \right] + \left[\frac{16}{3} \right]}{w(s) \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right]} \quad (2.37)$$

with

$$c(\infty) = -w(\infty) + \frac{16}{3} < \infty \quad (2.38)$$

verifying that the resultant compensator is, indeed, proper given a strictly proper plant.

Now, let us repeat the above example using the alternative coprime fractional representation of Equation 1.14 and 1.15 which yields

$$\begin{aligned} c(s) &= \frac{-w(s) \left[\frac{(s-2)(s+1)}{(s+2)^2} \right] + \left[\frac{16(s+2)}{3(s+1)} \right]}{w(s) \left[\frac{(s+1)^2}{(s+2)^3} \right] + \left[\frac{(s+2/3)}{(s+1)} \right]} \\ &= \frac{-w(s) \left[\frac{(s+1)^2}{(s+2)^2} \right] \left[\frac{(s-2)}{(s+2)} \right] + \left[\frac{16}{3} \right]}{w(s) \left[\frac{(s+1)^2}{(s+2)^2} \right] \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right]} \end{aligned}$$

FEEDBACK SYSTEM DESIGN

$$= \frac{-w'(s) \left[\frac{(s-2)}{(s+2)} \right] + \left[\frac{16}{3} \right]}{w'(s) \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right]} \quad (2.39)$$

where

$$w'(s) = w(s) \left[\frac{(s+1)^2}{(s+2)^2} \right] \quad (2.40)$$

As such, the same set of compensators are obtained from the alternative coprime fractional representation as from the original representation though the parameterizations defined differ by the miniphase factor $[(s+1)^2]/[(s+2)^2]$.

Finally, let us set $w'(s) = 0$ in 2.39 obtaining the compensator

$$c(s) = \frac{16(s+2)}{3(s+2/3)} \quad (2.41)$$

Now, $c(s)$ has a zero at $s = -2$ which cancels the pole of $p(s)$ at $s = -2$. This does not, however, contradict the requirement that $p(s)c(s) = [n_p(s)n_c(s)]/[d_p(s)d_c(s)]$ be coprime since our coprimeness concept only forbids right half-plane pole-zero cancellations. Of course, a left half-plane pole-zero cancellation such as encountered in the above example is benign and need not be forbidden.

3. Tracking

Once we have stabilized our feedback system we may use the remaining design latitude, the choice of a stable $w(s)$, to meet various system design constraints. The first of several such design constraints which we will consider are the asymptotic tracking and disturbance rejection conditions wherein we require that the system asymptotically follow or reject a prescribed input [4].

In the *tracking (or asymptotic regulator) problem* it is desired to design a stable feedback system whose output, v_2 , asymptotically follows a prescribed input which we model by the impulse response of a transfer function $t(s)$, as illustrated in Figure 2. As usual we assume that $t(s)$ admits a

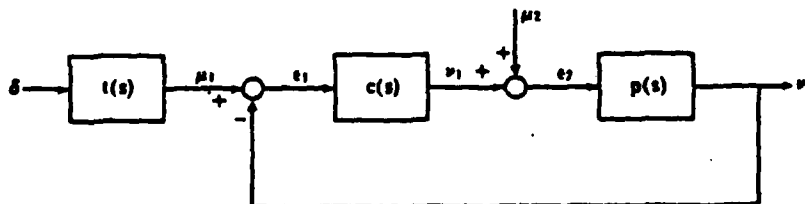


Figure 2. Feedback system with tracking generator.

coprime fractional representation in the form

$$t(s) = \frac{n_t(s)}{d_t(s)} \quad (3.1)$$

where there exist stable $u_t(s)$ and $v_t(s)$ such that

$$u_t(s)n_t(s) + v_t(s)d_t(s) = 1 \quad (3.2)$$

We say that the system tracks the impulse response of $t(s)$ if

$$t(s) - h_{n_p n_t}(s)t(s) = h_{n_1 n_t}(s)t(s) \quad (3.3)$$

is stable. Recall that the impulse response of a single-variate system is asymptotic to zero if and only if the corresponding transfer function is stable. Thus the response of our system to the impulse response of $t(s)$ will be asymptotic to the impulse response of $t(s)$ if and only if the transfer of equation 3.2 is stable.

Recall from Corollary 6 that

$$h_{n_1 n_t}(s) = w(s)n_p(s)d_p(s) + v_p(s)d_p(s) \quad (3.4)$$

hence if we desire to stabilize the system and simultaneously cause it to track the impulse response of $t(s)$ we must choose a stable $w(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero and

$$h_{n_1 n_t}(s)t(s) = \frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)]\tilde{n}_t(s)}{d_t(s)} \quad (3.5)$$

is stable.

11. Property. Given $p(s)$ there exists a compensator for the feedback system of Figure 2 which stabilizes the system and simultaneously causes it to track the impulse response of $t(s)$ if and only if the equation

$$w(s)n_p(s)d_p(s) + x(s)d_t(s) = u_p(s)n_p(s) - 1$$

admits stable solutions $w(s)$ and $x(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. In this case the required compensator takes the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where $w(s)$ is a solution to the above equation.

Proof. If there exists a stable $w(s)$ such that 3.5 is stable then it follows from 3.2 that

$$\frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)]}{d_t(s)}$$

FEEDBACK SYSTEM DESIGN

$$= \frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)] [u_i(s)n_i(s) + v_i(s)d_i(s)]}{d_i(s)} \quad (3.6)$$

$$= \left[\frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)] n_i(s)}{d_i(s)} \right] u_i(s) + [w(s)n_p(s)d_p(s) + v_p(s)d_p(s)] v_i(s) \\ = -x(s)$$

is stable since it is expressed as the sum of products of stable functions. Rearranging 3.6 and invoking 2.15 then yields

$$w(s)n_p(s)d_p(s) + x(s)d_i(s) = -v_p(s)d_p(s) = u_p(s)n_p(s) - 1 \quad (3.7)$$

as required. Conversely, if 3.7 admits stable solutions, $w(s)$ and $x(s)$, where $w(s)n_p(s) + v_p(s)$ is not identically zero we define $c(s)$ by

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]} \quad (3.8)$$

using the $w(s)$ of 3.7. Now, with this $w(s)$ 3.7 and 3.5 imply that

$$h_{e_1, n_1}(s) t(s) = \frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)] n_i(s)}{d_i(s)} \\ = \frac{[-x(s)d_i(s)] n_i(s)}{d_i(s)} = -x(s)n_i(s) \quad (3.9)$$

is stable. Since the stabilization theorem implies that any $c(s)$ in the form of 3.8 stabilizes the system while 3.9 implies that $h_{e_1, n_1}(s) t(s)$ is stable for this choice of $w(s)$ we have constructed the desired compensator.

Tracking Theorem: Given $p(s)$ there exists a compensator for the feedback system of Figure 2 which stabilizes the system and simultaneously causes it to track the impulse response of $t(s)$ if and only if $n_p(s)$ and $d_i(s)$ are coprime. In this case let $u_{pi}(s)$ and $v_{pi}(s)$ be stable functions such that

$$u_{pi}(s)n_p(s) + v_{pi}(s)d_i(s) = 1$$

and let $a(s) = n_p(s)/d_p(s)$ be a coprime fractional representation of $a(s) = d_p(s)/d_i(s)$. Then the desired set of compensators take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where

$$w(s) = -u_{pi}(s)v_p(s) + e(s)d_p(s)$$

with $e(s)$ an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero.

Proof. Consistent with Property 11, it suffices to show that the equation

$$w(s)n_p(s)d_p(s) + x(s)d_i(s) = u_p(s)n_p(s) - 1 \quad (3.10)$$

admits stable solutions $w(s)$ and $x(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero if and only if $n_p(s)$ and $d_i(s)$ are coprime. In this case we then show that the appropriate $w(s)$ takes the form

$$w(s) = -u_p(s)v_p(s) + e(s)d_p(s) \quad (3.11)$$

for any stable $e(s)$.

If 3.10 admits stable solutions $w(s)$ and $x(s)$ then it follows from 3.10 that

$$u_p(s)n_p(s) - w(s)n_p(s)d_p(s) - x(s)d_i(s) = 1 \quad (3.12)$$

or equivalently

$$[u_p(s) - w(s)d_p(s)]n_p(s) + [-x(s)]d_i(s) = 1 \quad (3.13)$$

showing that $n_p(s)$ and $d_i(s)$ are coprime. Conversely, if $n_p(s)$ and $d_i(s)$ are coprime there exists $u_p(s)$ and $v_p(s)$ such that

$$u_p(s)n_p(s) + v_p(s)d_i(s) = 1 \quad (3.14)$$

from which it follows that

$$\begin{aligned} u_p(s)n_p(s) - 1 &= -v_p(s)d_i(s) \\ &= -[u_p(s)n_p(s) + v_p(s)d_i(s)]v_p(s)d_p(s) \\ &= [-u_p(s)v_p(s)]n_p(s)d_p(s) + [-v_p(s)v_p(s)d_p(s)]d_i(s) \end{aligned} \quad (3.15)$$

As such,

$$w^p(s) = -u_p(s)v_p(s) \quad (3.16)$$

and

$$x^p(s) = -v_p(s)v_p(s)d_p(s) \quad (3.17)$$

represent particular solutions to 3.10.

To construct homogeneous solutions 3.10 we define a transfer function $a(s)$ by $a(s) = d_p(s)/d_i(s)$ and let

$$a(s) = \frac{n_p(s)}{d_p(s)} \quad (3.18)$$

be a coprime fractional representation for $a(s)$. It then follows from Property 2 that there exists a stable $k(s)$ such that

FEEDBACK SYSTEM DESIGN

$$d_p(s) = n_g(s)k(s) \quad (3.19)$$

and

$$d_i(s) = d_g(s)k(s) \quad (3.20)$$

Thus if we define candidates for the homogeneous solution of 3.10 by

$$w^h(s) = e(s)d_g(s) \quad (3.21)$$

and

$$x^h(s) = -e(s)n_p(s)n_g(s) \quad (3.22)$$

where $e(s)$ is an arbitrary stable rational function we have

$$\begin{aligned} w^h(s)n_p(s)d_p(s) + x^h(s)d_i(s) \\ = e(s)d_g(s)n_p(s)n_g(s)k(s) - e(s)n_p(s)n_g(s)d_g(s)k(s) = 0 \end{aligned} \quad (3.23)$$

verifying that 3.21 and 3.22 are, indeed, homogeneous solutions.

To complete the solution of 3.10 we must show that all homogeneous solutions are of the form 3.21 and 3.22. To this end assume that $\underline{w}^h(s)$ and $\underline{x}^h(s)$ are stable and satisfy

$$\underline{w}^h(s)n_p(s)d_p(s) + \underline{x}^h(s)d_i(s) = 0 \quad (3.24)$$

and define $e(s)$ by

$$e(s) = \underline{w}^h(s)d_g(s) \quad (3.25)$$

Clearly,

$$\underline{w}^h(s) = e(s)d_g(s) \quad (3.26)$$

while it follows from 3.24 that

$$\underline{x}^h(s) = \frac{-\underline{w}^h(s)n_p(s)d_p(s)}{d_i(s)} = -\underline{w}^h(s)n_p(s) \quad (3.27)$$

$$= \frac{-\underline{w}^h(s)n_p(s)n_g(s)}{d_g(s)} = -e(s)n_p(s)n_g(s)$$

showing that $w^h(s)$ and $x^h(s)$ have the form of 3.21 and 3.22. It remains, however, to verify that $e(s)$ is stable for which purpose we invoke equations 3.14 and the coprimeness of $n_g(s)$ and $d_g(s)$ from which it follows that there exists stable $u_g(s)$ and $v_g(s)$ such that

$$u_g(s)n_g(s) + v_g(s)d_g(s) = 1 \quad (3.28)$$

As such,

$$\begin{aligned}
 e(s) &= \frac{w^h(s)}{d_e(s)} = \left[\frac{w^h(s)}{d_e(s)} \right] [u_e(s)n_e(s) + v_e(s)d_e(s)] \\
 &= \frac{w^h(s)u_e(s)n_e(s)}{d_e(s)} + w^h(s)v_e(s) \\
 &= w^h(s)u_e(s)a(s) + w^h(s)v_e(s) = \frac{w^h(s)u_e(s)d_p(s)}{d_i(s)} + w^h(s)v_e(s) \\
 &= \left[\frac{w^h(s)u_e(s)d_p(s)}{d_i(s)} \right] [u_p(s)n_p(s) + v_p(s)d_i(s)] + w^h(s)v_e(s) \\
 &= \frac{w^h(s)u_e(s)d_p(s)u_p(s)n_p(s)}{d_i(s)} + w^h(s)u_e(s)d_p(s)v_p(s) + w^h(s)v_e(s) \\
 &= -w^h(s)u_e(s)u_p(s) + w^h(s)u_e(s)d_p(s)v_p(s) + w^h(s)v_e(s)
 \end{aligned} \tag{3.29}$$

is stable since we have expressed it as the sum of products of stable functions. Note, the last equality in 3.29 follows from 3.24.

The solution space for Equation 3.10 thus takes the form

$$w(s) = -u_p(s)v_p(s) + e(s)d_e(s) \tag{3.30}$$

and

$$x(s) = -v_p(s)v_p(s)d_p(s) - e(s)n_p(s)n_e(s) \tag{3.31}$$

As such, Property 11 implies that if the $w(s)$ of Equation 3.30 is used to define a stabilizing compensator as per the stabilization theorem it will also cause the system to track the impulse response of $T(s)$. Of course, we must also assume that $e(s)$ is chosen so that $w(s)n_p(s) + v_p(s)$ is not identically zero. To complete our proof that the coprimeness of $n_p(s)$ and $d_i(s)$ is a sufficient condition for the solution of the tracking problem it thus suffices to show that there exists at least one choice of $e(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. From 3.30 it follows that

$$w(s)n_p(s) + v_p(s) = [-u_p(s)n_p(s) + 1]v_p(s) + e(s)d_e(s)n_p(s) \tag{3.32}$$

Now, $d_e(s)$ is not identically zero since it is the denominator of $a(s)$. As such, if $n_p(s)$ is not identically zero it follows from 3.32 that the value of $w(s)n_p(s) + v_p(s)$ is a non-trivial function of $e(s)$ and is therefore not identical to zero for all $e(s)$. On the other hand if $n_p(s) = 0$ then 2.15 implies that $v_p(s)$ is miniphase which, in turn, implies that $w(s)n_p(s) + v_p(s) = v_p(s)$ is not identically zero. Our proof is therefore complete.

Although the proof of our theorem is long, though elementary, the basic

result to the effect that a compensator exists which will simultaneously stabilize the system and cause it to track the impulse of $t(s)$ if and only if $n_p(s)$ and $d_i(s)$ are coprime is simple to check (no common right half-plane zeros). Moreover, the construction of the required compensator is simply a matter of substitution as per the following example.

12. Example. Consider the problem of designing a compensator for the plant of Examples 4 and 10 so that the system is stable and asymptotically tracks a step function. Recall that

$$p(s) = \left[\frac{(s+1)}{(s^2-4)} \right] = \frac{\left[\frac{(s+1)}{(s+2)^2} \right]}{\left[\frac{(s-2)}{(s+2)} \right]} = \frac{n_p(s)}{d_p(s)} \quad (3.33)$$

where

$$\frac{16}{3} \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right] \left[\frac{(s-2)}{(s+2)} \right] = u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (3.34)$$

While we take $t(s) = 1/s$ from which we obtain

$$t(s) = \frac{1}{s} = \frac{\left[\frac{1}{(s+2)} \right]}{\left[\frac{s}{(s+2)} \right]} = \frac{n_t(s)}{d_t(s)} \quad (3.35)$$

Now, $n_p(s)$ has no right half-plane zeros while $d_i(s)$ has a right half-plane zero at $s = 0$. As such, $n_p(s)$ and $d_i(s)$ are coprime and according to the theorem the desired compensator exists. Indeed,

$$\left[4 \right] \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{s}{(s+2)} \right] \left[\frac{s}{(s+2)} \right] = u_{pi}(s)n_p(s) + v_{pi}(s)d_i(s) = 1 \quad (3.36)$$

As a final step in the construction of our compensator we let

$$a(s) = \frac{d_p(s)}{d_i(s)} = \frac{\left[\frac{(s-2)}{(s+2)} \right]}{\left[\frac{s}{(s+2)} \right]} = \frac{n_a(s)}{d_a(s)} \quad (3.37)$$

which is coprime since

$$\left[-1 \right] \left[\frac{(s-2)}{(s+2)} \right] + \left[2 \right] \left[\frac{2}{(s+2)} \right] = u_a(s)n_a(s) + v_a(s)d_a(s) \quad (3.38)$$

It thus follows from the theorem that the desired $w(s)$ takes the form

$$w(s) = -u_p(s)v_p(s) + e(s)d_g(s) = \left[\frac{-4(s+2/3)}{(s+2)} \right] + \left[\frac{s}{(s+2)} \right] e(s) \quad (3.39)$$

where $e(s)$ is an arbitrary stable function. Substitution of any such $w(s)$ into

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]} \quad (3.40)$$

thus defines the required compensator so long as $e(s)$ is chosen so that $w(s)n_p(s) + v_p(s)$ is not identically zero. To verify our solution we may substitute $w(s)$ into the formula for $h_{1,p_1}(s)$ of Corollary 6 obtaining

$$h_{1,p_1}(s) = \left[\frac{s}{(s+2)^4} \right] [s(s-2)(s+2/3) + (s+1)(s-2)e(s)] \quad (3.41)$$

Since $h_{1,p_1}(s)$ has a zero at $s = 0$ the required tracking property may now be verified by the final value theorem.

Now let us consider an alternative problem where we are required to track $e^{2t} \cup (t)$. Here our tracking generator is defined by

$$r(s) = \left[\frac{1}{(s-2)} \right] = \frac{\left[\frac{1}{(s+2)} \right]}{\left[\frac{(s-2)}{(s+2)} \right]} = \frac{n_r(s)}{d_r(s)} \quad (3.42)$$

As before $n_p(s)$ and $d_r(s)$ are coprime since

$$\left[\frac{16}{3} \right] \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right] \left[\frac{(s-2)}{(s+2)} \right] = u_p(s)n_p(s) + v_p(s)d_r(s) = 1 \quad (3.43)$$

Finally, for this example we have

$$a(s) = \frac{d_p(s)}{d_r(s)} = \frac{\left[\frac{(s-2)}{(s+2)} \right]}{\left[\frac{(s-2)}{(s+2)} \right]} = \frac{1}{1} = \frac{n_a(s)}{d_a(s)} \quad (3.44)$$

where

$$[0][1] + [1][1] = u_a(s)n_a(s) + v_a(s)d_a(s) = 1 \quad (3.45)$$

Note that in this example $d_p(s)$ and $d_r(s)$ are not coprime since they have a common right half-plane zero at $s = 2$. For our purposes, however, all that is required is a coprime fractional representation for $a(s) = d_p(s)/d_r(s)$ as constructed above. Using these new values for $u_p(s)$ and $d_a(s)$ we obtain

$$w(s) = - \left[\frac{16(s+2/3)}{3(s+2)} \right] + e(s) \quad (3.46)$$

for any stable $e(s)$. This, in turn, yields

$$h_{e_1 \mu_1}(s) = \left[\frac{(s-2)}{9(s+2)^4} \right] [(9s^3 - 6s^2 - 20s - 8) + 9(s+1)(s+2)e(s)] \quad (3.47)$$

for which the zero at $s = 2$ indicates tracking. Note that every stable $w(s)$ is obtained for some $e(s)$ and hence all stabilizing compensators track $e^{2t} \cup (t)$ in this example.

4. Disturbance rejection

There are two alternative disturbance rejection problems which arise naturally in our feedback system theory. Figure 3a indicates the configuration for the *input disturbance rejection problem* [16] wherein we desire to design a compensator which simultaneously stabilizes the system and causes it to asymptotically reject the impulse response of $r(s)$, i.e., the response of the system to the impulse response of $r(s)$ should be asymptotic to zero.

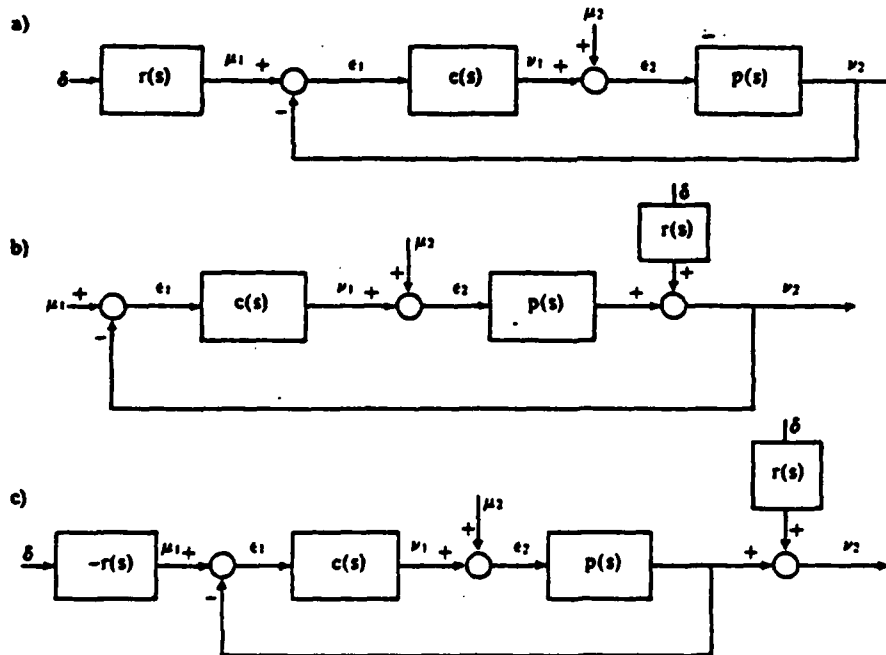


Figure 3. Feedback system configuration for a) the input disturbance rejection problem, b) the output disturbance rejection problem, and c) a modified configuration for the output disturbance rejection problem.

A similar *output disturbance rejection problem* [4] is illustrated in Figure 3b. Here, the disturbance is injected into the system at the plant output and, as before, it is desired to design a compensator which simultaneously stabilizes the system and causes it to asymptotically reject the impulse response of $r(s)$. Surprisingly, however, the output disturbance rejection problem is completely equivalent to the tracking problem considered in the previous section. To see this simply observe that the block diagram of Figure 3c is equivalent to that of Figure 3b. As such, if we design a compensator to stabilize the system and cause the plant output to asymptotically track the impulse response of $-r(s)$ when the impulse response of $r(s)$ is added to the plant output the total effect of the disturbance observed at y_2 will be asymptotic to zero. Consistent with the above we give no further consideration to the output disturbance rejection problem since it may be resolved via the techniques of the previous section with $t(s) = -r(s)$. Indeed, one can solve the tracking and output rejection problem simultaneously by working with tracking generator $t(s) = r(s)$.

For the input disturbance rejection problem we require that the impulse response of

$$h_{y_2 p_1}(s)r(s) = [-w(s)n_p(s)d_p(s) + u_p(s)n_p(s)]r(s) \quad (4.1)$$

be asymptotic to zero. Hence to simultaneously stabilize the feedback system and cause it to asymptotically reject the impulse response of $r(s)$ we must choose a stable $w(s)$ such that $w(s)n_p(s) + u_p(s)$ is not identically zero and $h_{y_2 p_1}(s)r(s)$ is stable. The required theory [10] is essentially identical to that used to solve the tracking problem and hence we simply state the pertinent theorems without proof. For this we let $r(s) = n_r(s)/d_r(s)$ be a coprime fractional representation for $r(s)$.

13. Property. Given $p(s)$ there exists a compensator for the feedback system of Figure 3a which stabilizes the system and simultaneously causes it to reject the impulse response of $r(s)$ if and only if the equation

$$w(s)n_p(s)d_p(s) + y(s)d_r(s) = u_p(s)n_p(s)$$

admits stable solutions $w(s)$ and $y(s)$ such that $w(s)n_p(s) + u_p(s)$ is not identically zero. In this case the required compensator takes the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + u_p(s)]}$$

where $w(s)$ is a solution to the above equation.

Disturbance Rejection Theorem. Given $p(s)$ there exists a compensator for the feedback system of Figure 3a which stabilizes the system and simultaneously causes it to reject the impulse response of $r(s)$ if and only if $d_p(s)$ and $d_r(s)$ are coprime. In this case let $u_{pr}(s)$ and $v_{pr}(s)$

be stable functions such that

$$u_{pr}(s)d_p(s) + v_{pr}(s)d_r(s) = 1$$

and let $b(s) = n_b(s)/d_b(s)$ be a coprime fractional representation of $b(s) = n_p(s)/d_r(s)$. Then the desired set of compensators take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where

$$w(s) = u_{pr}(s)u_p(s) + f(s)d_b(s)$$

with $f(s)$ an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero.

14. Example. Continuing with the plant of Example 12 let us consider the problem of designing a compensator to reject a step function, i.e., we let

$$r(s) = \begin{bmatrix} 1 \\ s \end{bmatrix} = \frac{\begin{bmatrix} 1 \\ (s+2) \end{bmatrix}}{\begin{bmatrix} s \\ (s+2) \end{bmatrix}} = \frac{n_r(s)}{d_r(s)}$$

Now,

$$\begin{bmatrix} -1 \\ 2 \end{bmatrix} \begin{bmatrix} (s-2) \\ (s+2) \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} s \\ (s+2) \end{bmatrix} = u_{pr}(s)d_p(s) + v_{pr}(s)d_r(s) = 1 \quad (4.3)$$

showing that $d_p(s)$ and $d_r(s)$ are coprime. Finally, we let

$$b(s) = \frac{n_p(s)}{d_r(s)} = \frac{\begin{bmatrix} (s+1) \\ (s+2)^2 \end{bmatrix}}{\begin{bmatrix} s \\ (s+2) \end{bmatrix}} = \frac{n_b(s)}{d_b(s)} \quad (4.4)$$

which is clearly coprime. From the theorem the required $w(s)$ take the form

$$w(s) = u_{pr}(s)u_p(s) + f(s)d_b(s) = \begin{bmatrix} -16 \\ 3 \end{bmatrix} + \begin{bmatrix} s \\ (s+2) \end{bmatrix} f(s) \quad (4.5)$$

where $f(s)$ is arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero. The use of the compensator derived from this $w(s)$ then results in the gains

$$h_{r_2u_1}(s) = \begin{bmatrix} s(s+1) \\ (s+2)^4 \end{bmatrix} \begin{bmatrix} 32(s+2) - (s-2)f(s) \\ 3 \end{bmatrix} \quad (4.6)$$

and

$$h_{1,p_1}(s) = \left[\frac{(s-2)}{3(s+2)^4} \right] [(s+2)(3s^2 - 8s - 12) + 3s(s+1)/(s)] \quad (4.7)$$

Here, the fact that $h_{1,p_1}(s)$ has a zero at $s = 0$ indicates that the disturbance rejection specification has been achieved while the fact that $h_{1,p_1}(s)$ has a zero at $s = 2$ implies that the system also tracks $e^{2t} \cup(t)$. This is consistent with Example 12 for the system tracks $e^{2t} \cup(t)$.

5. Simultaneous tracking and disturbance rejection

The purpose of this section is to combine the results of the previous two sections by formulating criteria for the design of a compensator which simultaneously stabilizes the system, causes it to track the impulse response of $t(s)$ and causes it to reject the impulse response of $r(s)$ [16]. The appropriate feedback system configuration is shown in Figure 4 where $r(s)$ is taken to be an input disturbance. Of course, an output disturbance can also be included in the theory simply by combining it with the tracking generator.

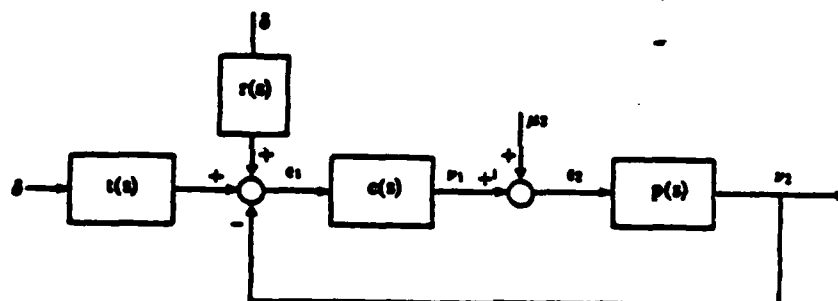


Figure 4. Configuration for the simultaneous tracking and disturbance rejection problem.

For consistency with the previous sections we will use the same notation which is reviewed as follows. Our plant is assumed to have a coprime fractional representation

$$p(s) = \frac{n_p(s)}{d_p(s)} \quad (5.1)$$

such that

$$u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (5.2)$$

while the tracking and rejection generators are characterized by coprime fractional representations

$$t(s) = \frac{n_t(s)}{d_t(s)} \quad (5.3)$$

and

$$r(s) = \frac{n_r(s)}{d_r(s)} \quad (5.4)$$

Also as in the previous sections we define $a(s)$ and $b(s)$ by

$$a(s) = \frac{d_p(s)}{d_i(s)} = \frac{n_o(s)}{d_o(s)} \quad (5.5)$$

and

$$b(s) = \frac{d_p(s)}{d_r(s)} = \frac{n_o(s)}{d_o(s)} \quad (5.6)$$

where $a(s) = n_o(s)/d_o(s)$ and $b(s) = n_b(s)/d_b(s)$ are coprime in the sense that there exists stable $u_o(s)$, $v_o(s)$, $u_b(s)$, and $v_b(s)$ such that

$$u_o(s)n_o(s) + v_o(s)d_o(s) = 1 \quad (5.7)$$

and

$$u_b(s)n_b(s) + v_b(s)d_b(s) = 1 \quad (5.8)$$

Moreover it follows from 5.5 and 5.6 together with property 2 that there exists stable $k(s)$ and $m(s)$ such that

$$d_i(s) = d_o(s)k(s) \text{ and } d_p(s) = n_o(s)k(s) \quad (5.9)$$

and

$$d_r(s) = d_b(s)m(s) \text{ and } n_p(s) = n_b(s)m(s) \quad (5.10)$$

Finally, the coprimeness conditions for the tracking and disturbance rejection problems are characterized by

$$u_{pr}(s)n_p(s) + v_{pr}(s)d_i(s) = 1 \quad (5.11)$$

and

$$u_{pr}(s)d_p(s) + v_{pr}(s)d_r(s) = 1 \quad (5.12)$$

while we will also require a coprimeness condition between $d_i(s)$ and $d_r(s)$ which we characterize by the equation

$$u_{ir}(s)d_i(s) + v_{ir}(s)d_r(s) = 1 \quad (5.13)$$

With this review of notation in hand the required design equations for the simultaneous tracking and disturbance rejection problem can be obtained simply by combining the results of Property 11 and Property 13 and observing that both design equations must be satisfied by the same $w(s)$ since we desire to construct a single compensator which simultaneously solves both problems.

15. Property. Given $p(s)$ there exists a compensator for the feedback system of Figure 4 which stabilizes the system, causes it to track the impulse response of $r(s)$ and simultaneously causes it to reject the impulse response of $r(s)$ if and only if the pair of equations

$$w(s)n_p(s)d_p(s) + x(s)d_r(s) = u_p(s)n_p(s) - 1$$

and

$$w(s)n_p(s)d_p(s) + y(s)d_r(s) = u_p(s)n_p(s)$$

admit stable solutions $w(s)$, $x(s)$, and $y(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. In this case the required compensator takes the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where $w(s)$ is a solution to the above equations.

Simultaneous Tracking and Disturbance Rejection Theorem: Given $p(s)$ there exists a compensator for the feedback system of Figure 4 which stabilizes the system, causes it to track the impulse response of $r(s)$ and simultaneously causes it to reject the impulse response of $r(s)$ if and only if

- (i) $n_p(s)$ and $d_r(s)$ are coprime,
- (ii) $d_p(s)$ and $d_r(s)$ are coprime, and
- (iii) $d_p(s)$ and $d_r(s)$ are coprime.

In that case the desired set of compensators take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where

$$w(s) = [u_p(s)u_p(s)u_r(s)d_r(s) - u_p(s)v_p(s)u_r(s)d_r(s)] + g(s)d_p(s)d_r(s)$$

with $g(s)$ an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero.

Proof. The fact that $n_p(s)$ and $d_r(s)$ must be coprime follows from the tracking theorem while the fact that $d_p(s)$ and $d_r(s)$ must be coprime follows from the disturbance rejection theorem. To verify that $d_p(s)$ and $d_r(s)$ must also be coprime for simultaneous stabilization we subtract the two design equations of Property 15 obtaining

$$[y(s)]d_r(s) + [-x(s)]d_r(s) = 1 \quad (5.14)$$

Conversely, to show that the three coprimeness conditions are also sufficient we must construct a $w(s)$ which simultaneously satisfies the criteria for tracking and disturbance rejections derived in the preceding sections. Upon invoking the results of the tracking and disturbance rejection

theorems we must therefore solve

$$-u_{pr}(s)v_p(s) + e(s)d_a(s) = w(s) = u_{pr}(s)u_p(s) + f(s)d_b(s) \quad (5.15)$$

for stable $e(s)$, $f(s)$, and $w(s)$. Since the required set of stable $w(s)$ may be obtained by substitution once $e(s)$ and $f(s)$ have been parameterized our main problem is to characterize the stable solutions of

$$e(s)d_a(s) - f(s)d_b(s) = [u_{pr}(s)u_p(s) + u_{pr}(s)v_p(s)] \quad (5.16)$$

To obtain a particular solution for 5.16 we invoke 5.9, 5.10, and 5.13 obtaining

$$\begin{aligned} & [u_{pr}(s)u_p(s) + u_{pr}(s)v_p(s)] \\ &= [u_{pr}(s)u_p(s) + u_{pr}(s)v_p(s)] [u_{ir}(s)d_i(s) + v_{ir}(s)d_r(s)] \\ &= [u_{pr}(s)u_p(s) + u_{pr}(s)v_p(s)] [u_{ir}(s)k(s)d_a(s) + v_{ir}(s)m(s)d_b(s)] \\ &= \{ [u_{pr}(s)u_p(s) + u_{pr}(s)v_p(s)] u_{ir}(s)k(s) \} d_a(s) \\ & \quad + \{ [u_{pr}(s)u_p(s) + u_{pr}(s)v_p(s)] v_{ir}(s)m(s) \} d_b(s) \end{aligned} \quad (5.17)$$

As such, the required particular solutions take the form

$$e^p(s) = [u_{pr}(s)u_p(s) + u_{pr}(s)v_p(s)] u_{ir}(s)k(s) \quad (5.18)$$

and

$$f^p(s) = - [u_{pr}(s)u_p(s) + u_{pr}(s)v_p(s)] v_{ir}(s)m(s) \quad (5.19)$$

Of course,

$$e^h(s) = g(s)d_b(s) \quad (5.20)$$

and

$$f^h(s) = g(s)d_a(s) \quad (5.21)$$

represent homogeneous solutions to 5.16 for any stable $g(s)$ since

$$e^h(s)d_a(s) - f^h(s)d_b(s) = g(s)d_b(s)d_a(s) - g(s)d_a(s)d_b(s) = 0 \quad (5.22)$$

As such, our characterization of the solution space for 5.16 will be complete if we can verify that all stable homogeneous solutions to 5.16 take the form of 5.20 and 5.21 for some stable $g(s)$. To this end let $\underline{e}^h(s)$ and $\underline{f}^h(s)$ be arbitrary stable homogeneous solutions to 5.16, i.e.,

$$\underline{e}^h(s)d_a(s) - \underline{f}^h(s)d_b(s) = 0 \quad (5.23)$$

Now, define $g(s)$ by $g(s) = \underline{e}^h(s)/d_b(s)$ in which case

$$\underline{e}^h(s) = g(s)d_b(s) \quad (5.24)$$

while 5.23 implies that

$$\underline{f}^h(s) = \frac{\underline{e}^h(s) d_g(s)}{d_b(s)} = g(s) d_g(s) \quad (5.25)$$

showing that $\underline{e}^h(s)$ and $\underline{f}^h(s)$ are of the required form. It remains, however, to show that $g(s)$ is stable. Indeed,

$$\begin{aligned} g(s) &= \frac{\underline{e}^h(s)}{d_b(s)} = \frac{\underline{e}^h(s)}{d_b(s)} [u_b(s) n_b(s) + v_b(s) d_b(s)] \\ &= \underline{e}^h(s) v_b(s) + \frac{\underline{e}^h(s) u_b(s) n_b(s)}{d_b(s)} \\ &= \underline{e}^h(s) v_b(s) + \frac{\underline{e}^h(s) u_b(s) n_p(s)}{d_r(s)} [u_r(s) d_r(s) + v_r(s) d_r(s)] \\ &= \underline{e}^h(s) v_b(s) + \underline{e}^h(s) u_b(s) n_p(s) v_r(s) + \frac{\underline{e}^h(s) u_b(s) n_p(s) u_r(s) d_r(s)}{d_r(s)} \\ &= \underline{e}^h(s) v_b(s) + \underline{e}^h(s) u_b(s) n_p(s) v_r(s) + \frac{\underline{e}^h(s) u_b(s) n_p(s) u_r(s) d_g(s) k(s)}{d_b(s)} \\ &= \underline{e}^h(s) v_b(s) + \underline{e}^h(s) u_b(s) n_p(s) v_r(s) + \underline{f}^h(s) u_b(s) n_p(s) u_r(s) k(s) \quad (5.26) \end{aligned}$$

showing that $g(s)$ is stable since it has been expressed as the sum of products of stable functions. Here, Equation 5.26 was derived with the aid of Equations 5.6, 5.8, 5.9, 5.13, and 5.23. As such, the complete set of solutions to 5.16 take the form

$$e(s) = [u_{pr}(s) u_p(s) + u_{pr}(s) v_p(s)] u_r(s) k(s) + g(s) d_b(s) \quad (5.27)$$

and

$$f(s) = [u_{pr}(s) u_p(s) + u_{pr}(s) v_p(s)] v_r(s) m(s) + g(s) d_g(s) \quad (5.28)$$

Now, upon substituting either of these expressions into 5.15 we obtain the desired expression for $w(s)$ in the form

$$\begin{aligned} w(s) &= -u_{pr}(s) v_p(s) + u_{pr}(s) u_p(s) + u_{pr}(s) v_p(s) u_r(s) k(s) d_g(s) + g(s) d_b(s) d_g(s) \\ &= u_{pr}(s) u_p(s) u_r(s) d_r(s) + u_{pr}(s) v_p(s) u_r(s) d_r(s) - 1 + g(s) d_b(s) d_g(s) \\ &= [u_{pr}(s) u_p(s) u_r(s) d_r(s) - u_{pr}(s) v_p(s) v_r(s) d_r(s)] + g(s) d_b(s) d_g(s) \quad (5.29) \end{aligned}$$

where $g(s)$ is an arbitrary function.

Finally, to complete the proof that the three coprimeness conditions suffice for simultaneous tracking and disturbance rejection we must verify that there exists one choice of a stable $g(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. Upon substituting 5.29 into this expression we obtain

$$\begin{aligned} w(s)n_p(s) + v_p(s) = & [u_{rr}(s)u_p(s)u_r(s)d_i(s) - u_{rr}(s)v_p(s)u_r(s)d_r(s)]n_p(s) \\ & + v_p(s) + g(s)d_b(s)d_o(s)n_p(s) \end{aligned} \quad (5.30)$$

Now $d_b(s)$ and $d_o(s)$ are not identically zero since they represent the denominators for well defined transfer functions. As such, if $n_p(s)$ is not identically zero 5.30 will be non-trivially dependent on $g(s)$ and hence not identically zero for every choice of $g(s)$. On the other hand if $n_p(s)$ is identically zero the equality $u_{rr}(s)n_p(s) + v_p(s)d_p(s) = 1$ implies that $v_p(s)$ is miniphase hence

$$w(s)n_p(s) + v_p(s) = v_p(s) \quad (5.31)$$

is not identically zero. Our proof is therefore complete.

Although the theorem is highly complex and, indeed, is predicated on the equally complex theorems which preceded it, the final result is an explicit description of the desired family of compensators. Moreover, the terms in this expression are readily computed by solving one or more coprimeness equation. As such, the result is easily implemented as per the following example.

16. Example. Continuing our analysis of the plant introduced in the previous examples we will investigate the possibility of simultaneously tracking $e^{2t}U(t)$ and rejecting $U(t)$. Here,

$$d_i(s) = \frac{(s-2)}{(s+2)} \quad (5.32)$$

and

$$d_r(s) = \frac{s}{(s+2)} \quad (5.33)$$

which are clearly coprime. Indeed

$$\begin{bmatrix} -1 \end{bmatrix} \begin{bmatrix} \frac{(s-2)}{(s+2)} \end{bmatrix} + \begin{bmatrix} 2 \end{bmatrix} \begin{bmatrix} \frac{s}{(s+2)} \end{bmatrix} = u_{rr}(s)d_i(s) + v_{rr}(s)d_r(s) = 1 \quad (5.34)$$

As such, the required $w(s)$ takes the form

$$w(s) = \left[\frac{1}{(s+2)^2} \right] \left[\frac{-16}{3} (s^2 + \frac{4}{3}s + 4) + s(s+2)g(s) \right] \quad (5.35)$$

yielding

$$h_{2,1}(s) = \left[\frac{(s+1)s}{(s+2)^5} \right] \left[\frac{32}{3}(s^2 + \frac{8}{3}s + \frac{20}{3}) - (s-2)(s+2)g(s) \right] \quad (5.36)$$

and

$$h_{1,1}(s) = \left[\frac{(s-2)}{9(s+2)^5} \right] \left[(9s^4 + 12s^3 + 32s^2 - 112s - 144) + 9s(s+1)(s+2)g(s) \right] \quad (5.37)$$

which have the required zeros at $s = 0$ and $s = 2$, respectively.

References

1. Antsaklis, P.J., and J.B. Pearson, "Stabilization and Regulation in Linear Multivariate Systems", *IEEE Trans. on Auto. Cont.*, Vol. AC-23, pp. 928-930, (1978).
2. Barnett, S. *Matrices in Control Theory*, London, Van Nostrand, 1971.
3. Bengtsson, G., "Output Regulation and Internal Modes - A Frequency Domain Approach", *Automatica*, Vol. 13, pp. 333-345, (1977).
4. Callier, F.M., and C.A. Desoer, "Stabilization, Tracking and Disturbance Rejection in Linear Multivariable Distributed Systems" *Proc. of the 17th IEEE Conf. on Decision and Control*, San Diego, Jan. 1979, pp. 513-514, (also Tech Memo UCB/ERL M78/83, Univ. of California at Berkeley, Dec. 1978).
5. Chang, L., and J.B. Pearson, "Frequency Domain Synthesis of Multivariable Linear Regulators", *IEEE Trans. on Auto. Cont.*, Vol. AC-23, pp. 3-15, (1978).
6. Cheng, L., and J.B. Pearson, "Synthesis of Linear Multivariable Regulators", *IEEE Trans. on Auto. Cont.*, (to appear).
7. Chua, O., M.S. Thesis, Texas Tech Univ., 1980.
8. Desoer, C.A., Liu, R.-w., Murray, J., and R. Sacks, "Feedback System Design: the fractional representation approach to analysis and synthesis", *IEEE Trans. on Auto. Cont.*, Vol. AC-25, pp. 399-412, (1980).
9. Feintuch, A., and R. Sacks, *System Theory: A Hilbert Space Approach*, New York, Academic Press, (to appear).
10. Francis, B., "The Multivariable Servomechanism Problem from the Input-Output Viewpoint", *IEEE Trans. on Auto. Cont.*, Vol. AC-22, pp. 322-328, (1977).
11. Francis, B., and M. Vidyasagar, "Algebraic and Topological Aspects of the Servo Problem for Lumped Linear Systems", unpublished notes, Yale Univ., 1980.
12. Heiton, J.W., "Orbit Structure of the Moebius Transformation Semigroup Acting on H^{∞} " in *Topics in Functional Analysis*, Advances in Mat. Supp. Studies, Vol. 3, New York, Academic Press, 1978, pp. 129-157.
13. Pernebo, L., Ph.D. Thesis, Lund Inst. of Tech., 1978.
14. Pernebo, L., "An Algebraic Theory for Design of Controllers for Linear Multivariable Systems; Parts I and II", *IEEE Trans. on Auto. Cont.*, (to appear).

FEEDBACK SYSTEM DESIGN

15. Rosenbrock, H.H., *State-Space and Multivariable Theory*, New York, J. Wiley and Sons, 1970.
16. Sacks, R., and J. Murray, "Feedback Systems Design: the tracking and disturbance rejection problems", *IEEE Trans. on Auto. Cont.*, Vol. AC-26, pp. 203-217, (1981).
17. Sacks, R., and J. Murray, "Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem", unpublished notes, Texas Tech Univ., 1980.
18. Vidyasagar, M., Schneider, H., and B. Francis, "Algebraic and Topological Aspects of Feedback System Stabilization", Tech Rpt. 80-90, Dept. of Elec. Engrg., Univ. of Waterloo, 1980.
19. Vidyasagar, M., and M. Viswanadham, "Algebraic Design Techniques for Reliable Stabilization", Tech Rpt. 81-02, Dept. of Elec. Engrg., Univ. of Waterloo, 1980.
20. Wolovitch, W.A., "Multivariable System Synthesis with Step Disturbance Rejection", *IEEE Trans. on Auto. Cont.*, Vol. AC-19, pp. 127-130, (1974).
21. Wolovitch, W.A., and P. Ferreira, "Output Regulation and Tracking in Linear Multivariable Systems", *IEEE Trans. on Auto. Cont.*, Vol. AC-24, pp. 460-465, (1979).
22. Youla, D.C., "Interpolary Multichannel Spectral Estimation", Unpublished Notes, Polytechnic Inst. of New York (1979).
23. Youla, D.C., Bongiorno, J.J., and H.A. Jabr, "Modern Wiener-Hopf Design of Optimal Controllers, Part I", *IEEE Trans. on Auto. Cont.*, Vol. AC-21, pp. 3-15, (1976).
24. Youla, D.C., Bongiorno, J.J., and H.A. Jabr, "Modern Wiener-Hopf Design of Optimal Controllers, Part II", *IEEE Trans. on Auto. Cont.*, Vol. AC-21, pp. 319-338, 1976.
25. Youla, D.C., Bongiorno, J.J., and C.N. Lu, "Single-Loop Feedback Stabilization of Linear Multivariable Dynamical Plants", *Automatica*, Vol. 10, pp. 159-173, (1974).
26. Zames, G., "Feedback and Optimal Sensitivity: Model reference transformation, weighted seminorms, and approximate inverses", *IEEE Trans. on Auto. Cont.*, (to appear).

FEEDBACK SYSTEM DESIGN:

The Single-Variate Case – Part II*

*R. Saeks¹, J. Murray¹, O. Chua², C. Karmokolias³,
and A. Iyer¹*

Summary of Part I. The present article represents a survey of a new frequency domain approach to the feedback system design problem. Although the theory applies to the general multivariate case and, in fact, much of the theory can be extended to a general ring theoretic setting for the purpose of the present exposition we will restrict ourselves to the single-variate case wherein the simple algebraic nature of the new theory is most readily apparent.

In Part I¹ we introduced the concept of a stable rational fractional representation for a feedback system and derived an asymptotic design theory therefrom. This included a complete parameterization for the set of compensators which stabilize the system under appropriate tracking and/or disturbance rejection constraints. Moreover, expressions for the feedback system gains of the resulting system which are linear (actually affine) in the underlying design parameter are obtained. As such, the process of choosing a specific design within this set to meet additional specifications is greatly simplified.

In Part II we investigate a number of these constraints. These include the robust design problem, transfer function design, the pole placement problem, the simultaneous design problem, and the problem of designing a stable stabilizing compensator.

6. Robust design

Although the design algorithms formulated in the preceding sections theoretically achieve the prescribed specification, in practice, they are often inapplicable since the specified plant model may not be exact. As such, we would like to formulate a design theory which is *robust* in the sense that the

* Received September 18, 1981; revised October 20, 1981. This research supported in part by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

¹ Department of Electrical Engineering, Texas Tech University, Lubbock, Texas 79409, USA.

² Presently with Honeywell, Inc., Phoenix, Arizona.

³ Presently with Kearfott Division, Singer, Inc., Little Falls, New Jersey.

prescribed design specifications are achieved for any plant in a "neighborhood" of the nominal plant $p(s)$. Then, if $p(s)$ is a "reasonable" approximation to the actual plant, our design formulated in terms of $p(s)$ will also work for the actual plant. Although, in practice, $p(s)$ is not an exact model for the plant it is reasonable to assume that $c(s)$, $r(s)$, and $t(s)$ are exact. Indeed, $c(s)$ is under control of the system designer while $t(s)$ and $r(s)$ are simply mathematical artifices which model the signals we desire to track and/or reject.

In various abstract control theories one can spend hours making the term "small-perturbation" precise [18]. For our single-variate case, however, it suffices to say that $\bar{p}(s) = n_{\bar{p}}(s)/d_{\bar{p}}(s)$ is close to $p(s) = n_p(s)/d_p(s)$ if the coefficients of $n_{\bar{p}}(s)$ are close to those of $n_p(s)$ and those of $d_{\bar{p}}(s)$ are close to those of $d_p(s)$. Indeed, it can be shown that this is a well defined concept so long as both fractional representations are coprime [11,18]. Since the poles and zeros of a rational function are continuous functions of its coefficients it then follows that a small perturbation of a miniphase function is miniphase [18]. That is, since both the poles and zeros of such a function lie in the (strict) left half-plane a small perturbation of its coefficients will not move them to the right half-plane.

The above observation, however, is sufficient to prove that any stabilizing controller is robust. Indeed, our stabilizing controllers are designed so that

$$d_p(s)d_c(s) + n_p(s)n_c(s) = 1 \quad (6.1)$$

As such, if the nominal plant, $p(s) = n_p(s)/d_p(s)$, is replaced by the actual plant $\bar{p}(s) = n_{\bar{p}}(s)/d_{\bar{p}}(s)$ but we still use our design compensator, the common denominator, for our feedback systems gains will become

$$d_{\bar{p}}(s)d_c(s) + n_{\bar{p}}(s)n_c(s) = k(s) \quad (6.2)$$

which is still miniphase if $\bar{p}(s)$ is close to $p(s)$. Thus the feedback system will still be stable even though the common denominator is no longer the identity.

17. Property. Every stabilizing compensator is robust.

Unlike the design of a simple stabilizing compensator wherein robustness takes care of itself, a compensator which meets a tracking and/or disturbance rejection specification may fail to be robust [11]. Interestingly, however, in our single-variate case if the coprimeness conditions required for the existence of a solution to the tracking and/or disturbance rejection problem are satisfied then a robust solution is guaranteed to exist though the set of robust compensators may be a strict subset of the set of compensators formulated in the previous sections.

Recall that the solution of the tracking problem requires a stabilizing compensator such that $h_{e,\mu_1}(s)t(s)$ is also stabilized. Now, if we use the

actual plant, $\bar{p}(s) = \bar{n}_p(s) / \bar{d}_p(s)$, rather than the nominal plant Equations 2.9 and 3.1 yield

$$\bar{h}_{\epsilon, \mu_1}(s) t(s) = \frac{\bar{d}_p(s) d_c(s) n_t(s)}{[d_p(s) d_c(s) + n_p(s) n_c(s)] d_t(s)} \quad (6.3)$$

The robust stabilization property implies that $d_p(s) d_c(s) + n_p(s) n_c(s)$ is miniphase when $\bar{p}(s)$ is close to $p(s)$ and hence 6.3 will be stable if and only if the numerator factors in 6.3 cancel any instabilities in $d_t(s)$. Of course, $n_t(s)$ and $d_t(s)$ are coprime implying that $n_t(s)$ will not help to cancel the instabilities in $d_t(s)$. On the other hand we cannot use $\bar{d}_p(s)$ to cancel these instabilities since it is not known exactly². As such, in a robust compensator $d_c(s)$ must cancel any instabilities in $d_t(s)$. In our setting this reduces to the requirement that

$$d_c(s) = -x'(s) d_t(s) \quad (6.4)$$

for some stable $x'(s)$. Here, the minus sign has been assumed for notational convenience only. Of course, it follows from the stabilization theorem that in every stabilizing compensator $d_c(s)$ takes that form

$$d_c(s) = w(s) n_p(s) + v_p(s) \quad (6.5)$$

Upon combining 6.4 and 6.5 we thus arrive at the following design equation for the desired robust tracker.

18. Property. Given $p(s)$ there exists a compensator for the feedback system of Figure 2 which robustly stabilizes the system and simultaneously causes it to robustly track the impulse response of $t(s)$ if and only if the equation

$$w(s) n_p(s) + x'(s) d_t(s) = -v_p(s)$$

admits stable solutions $w(s)$ and $x'(s)$ such that $w(s) n_p(s) + v_p(s)$ is not identically zero. In this case the required compensator takes the form

$$c(s) = \frac{[-w(s) d_p(s) + u_p(s)]}{[w(s) n_p(s) + v_p(s)]}$$

where $w(s)$ is a solution of the above equation.

Note that if we multiply the above design equation by $d_p(s)$ we obtain

$$w(s) n_p(s) d_p(s) + [x'(s) d_p(s)] d_t(s) = -v_p(s) d_p(s) = u_p(s) n_p(s) - 1 \quad (6.6)$$

hence if $w(s)$ satisfies the above design equation it also satisfies the design equation of Property 11 with $x(s) = x'(s) d_p(s)$. As such, our compensators for the robust tracking problem do, indeed, satisfy the criteria developed in Section 3 for the tracking problem. On the other hand some of the $w(s)$'s which satisfy the design equation of Property 11 may fail to satisfy the above design equation.

To parameterize the solution space for the robust tracking problem it suffices to characterize the stable solutions of

$$w(s)n_p(s) - x'(s)d_i(s) = -v_p(s) \quad (6.7)$$

Indeed, the same coprimeness condition employed in the tracking theorem also suffices for robust tracking even though the resultant solution space is smaller.

Robust Tracking Theorem: Given $p(s)$ there exists a compensator for the feedback system of Figure 2 which robustly stabilizes the system and simultaneously causes it to robustly track the impulse response of $t(s)$ if and only if $n_p(s)$ and $d_i(s)$ are coprime. In this case let $u_{pi}(s)$ and $v_{pi}(s)$ be stable functions such that

$$u_{pi}(s)n_p(s) + v_{pi}(s)d_i(s) = 1$$

Then the desired set of compensators takes the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where

$$w(s) = -u_{pi}(s)v_p(s) + e'(s)d_i(s)$$

with $e'(s)$ an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero.

Proof. Since any robust tracker is also a tracker the necessity for the coprimeness condition follows from the tracking theorem. To show that the coprimeness condition is also sufficient we multiply the equality

$$u_{pi}(s)n_p(s) + v_{pi}(s)d_i(s) = 1 \quad (6.8)$$

by $-v_p(s)$ obtaining a particular solution to 6.7 in the form

$$w^p(s) = -u_{pi}(s)v_p(s) \quad (6.9)$$

and

$$x'^p(s) = -v_{pi}(s)v_p(s) \quad (6.10)$$

Similarly, the homogeneous solutions to 6.7 take the form

$$w^h(s) = e'(s)d_i(s) \quad (6.11)$$

and

$$x'^h(s) = -e'(s)n_p(s) \quad (6.12)$$

As in the previous derivations one can use the coprimeness of $n_p(s)$ and $d_i(s)$ to verify that 6.11 and 6.12 range over all possible homogeneous solutions of 6.7 as $e'(s)$ ranges over the stable functions. The set of $w(s)$ which

achieves robust tracking thus takes the form

$$w(s) = -v_p(s)v_p(s) + e'(s)d_i(s) \quad (6.13)$$

where $e'(s)$ is an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero. Also as in the previous derivations one can show that a stable $e'(s)$ exists such that $w(s)n_p(s) + v_p(s)$ is not identically zero which completes the argument that the coprimeness condition is sufficient for robust tracking.

If one recalls from equation 5.9 that $d_i(s) = k(s)d_p(s)$ for some stable $k(s)$ and compares 6.13 with the corresponding equation for the tracking problem we conclude that the robust trackers correspond to the subset of all trackers defined by a stable function, $e(s)$, which takes the form

$$e(s) = k(s)e'(s) \quad (6.14)$$

As such, we can construct a family of robust trackers simply by using this class of $e(s)$ in the formulae derived in Section 3. Moreover, since $d(s)$ will be a unit if and only if $d_p(s)$ and $d_i(s)$ are coprime we have the following corollary.

19. Corollary. Every solution to the tracking problem is robust if and only if $d_p(s)$ and $d_i(s)$ are coprime.

It is also interesting to note that the derivation of our robust tracker is considerably simpler than that of the (non-robust) tracker since we do have to construct a coprime fractional representation for $a(s) = d_p(s)/d_i(s)$. Indeed, since small perturbations on $p(s)$ are assumed $\tilde{d}_p(s)$ and $\tilde{d}_i(s)$ will always be coprime for some choices of $\tilde{p}(s)$ and, as such, the robust theory cannot exploit cancellations between the nominal $d_p(s)$ and $d_i(s)$ even though they may not be coprime.

20. Example. Consistent with Equation 6.14 one can construct robust trackers for the plant

$$p(s) = \left[\frac{(s+1)}{(s^2-4)} \right] \quad (6.15)$$

of Example 12 simply by letting $e(s) = k(s)e'(s)$ in those examples. In the case where we desire to track a step function $d_p(s)$ and $d_i(s)$ are coprime via Equation 3.38 and, as such, the corollary implies that the entire set of tracking compensators characterized by Equations 3.39 and 3.40 are robust.

On the other hand if we desire to track $e^{2t} \cup (t)$, $d_p(s)$ and $d_i(s)$ are no longer coprime as per Equation 3.44. Indeed, in this case

$$d_i(s) = \left[\frac{(s-2)}{(s+2)} \right] d_p(s) = k(s)d_p(s) \quad (6.16)$$

and, as such, the substitution of 6.14 into 3.46 yields a set of robust compensators characterized by

$$w(s) = - \left[\frac{16(s+2/3)}{3(s+2)} \right] + \left[\frac{(s-2)}{(s+2)} \right] e'(s) \quad (6.17)$$

where $e'(s)$ is an arbitrary stable function.

The arguments required to design compensators for robust disturbance rejection and/or simultaneous robust tracking and robust disturbance rejection follow immediately upon combining the techniques used in Sections 4 and 5 with the ideas discussed above. As such, we will simply sketch the ideas and state the main results. For robust disturbance rejection we require that

$$\bar{h}_{v_2 u_1}(s) r(s) = \frac{n_{\bar{p}}(s) n_c(s) n_r(s)}{d_{\bar{p}}(s) d_c(s) + n_{\bar{p}}(s) n_c(s) d_r(s)} \quad (6.18)$$

be stable for all $\bar{p}(s)$ in a neighborhood of $p(s)$. As with the tracking problem this will be achieved over the entire neighborhood if and only if there exists a stable $y'(s)$ such that $n_c(s) = y'(s) d_r(s)$ in which case

$$n_c(s) = -w(s) d_p(s) + u_p(s) = y'(s) d_r(s) \quad (6.19)$$

21. Property. Given $p(s)$ there exists a compensator for the feedback system of Figure 3a which robustly stabilizes the system and simultaneously causes it to robustly reject the impulse response of $r(s)$ if and only if the equation

$$w(s) d_p(s) + y'(s) d_r(s) = u_p(s)$$

admits stable solutions $w(s)$ and $y'(s)$ such that $w(s) n_p(s) + v_p(s)$ is not identically zero. In this case the required compensator takes the form

$$c(s) = \frac{[-w(s) d_p(s) + u_p(s)]}{[w(s) n_p(s) + v_p(s)]}$$

where $w(s)$ is a solution of the above equation.

Robust Disturbance Rejection Theorem: Given $p(s)$ there exists a compensator for the feedback system of 3a which robustly stabilizes the system, and simultaneously causes it to robustly reject the impulse response of $r(s)$ if and only if $d_p(s)$ and $d_r(s)$ are coprime. In this case let $u_{pr}(s)$ and $v_{pr}(s)$ be stable functions such that

$$u_{pr}(s) d_p(s) + v_{pr}(s) d_r(s) = 1$$

Then the desired compensators take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where

$$w(s) = u_{pr}(s)u_p(s) + f'(s)d_r(s)$$

with $f'(s)$ an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero.

22. Corollary. Every solution to the disturbance rejection problem is robust if and only if $n_p(s)$ and $d_r(s)$ are coprime.

23. Corollary. Given $p(s)$ there exists a compensator for the feedback system of Figure 4 which robustly stabilizes the system, causes it to robustly track the impulse response of $t(s)$ and simultaneously causes it to robustly reject the impulse to $r(s)$ if and only if the pair of equations

$$w(s)n_p(s) + x'(s)d_r(s) = -v_p(s)$$

and

$$w(s)d_p(s) + y'(s)d_r(s) = u_p(s)$$

admits stable solutions $w(s)$, $x(s)$, and $y(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. In this case the required compensators take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where $w(s)$ is a solution to the above equations.

Simultaneous Robust Tracking and Robust Disturbance Rejection Theorem: Given $p(s)$ there exists a compensator for the feedback system of Figure 4 which robustly stabilizes the system, causes it to robustly track the impulse response of $t(s)$ and simultaneously causes it to robustly reject the impulse response of $r(s)$ if and only if

- (i) $n_p(s)$ and $d_r(s)$ are coprime,
- (ii) $d_p(s)$ and $d_r(s)$ are coprime, and
- (iii) $d_r(s)$ and $d_r(s)$ are coprime.

In that case the desired set of compensators take the form

$$c(s) = \frac{-w(s)d_p(s) + u_p(s)}{w(s)n_p(s) + v_p(s)}$$

where

$$w(s) = [u_{pr}(s)u_p(s)u_{rr}(s)d_r(s) - u_{pr}(s)v_p(s)v_{rr}(s)d_r(s)] + g'(s)d_r(s)d_r(s)$$

with $g(s)$ an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero.

Corollary. Every solution to the simultaneous tracking and disturbance rejection problem is robust if and only if

- (i) $d_p(s)$ and $d_i(s)$ are coprime and
- (ii) $n_p(s)$ and $d_i(s)$ are coprime.

7. Transfer function design

Rather than asking that a system meet a tracking or disturbance rejection specification we may attempt to fully specify one or more of the system gains while simultaneously stabilizing the system. For instance we may require that

$$h_{r_2 u_1}(s) = h(s) \quad (7.1)$$

where $h(s)$ is a prescribed stable transfer function. Of course Corollary 2.6 implies that

$$h_{r_2 u_1}(s) = -w(s)n_p(s)d_p(s) + u_p(s)n_p(s) \quad (7.2)$$

so that our stable design problem reduces to finding a stable $w(s)$ such that

$$h(s) = -w(s)n_p(s)d_p(s) + u_p(s)n_p(s) \quad (7.3)$$

Transfer Function Design Theorem: Given $p(s) \neq 0$ there exists a compensator for the feedback system of Figure 1 which stabilizes the system and yields $h_{r_2 u_1}(s) = h(s)$ for some stable $h(s)$ if and only if

- (i) $n_p(s)$ divides $h(s)$ and
- (ii) $d_p(s)$ divides $[1 - h(s)]$.

Proof. If 7.3 is satisfied

$$h(s) = -w(s)n_p(s)d_p(s) + u_p(s)n_p(s) = [-w(s)d_p(s) + u_p(s)]n_p(s) \quad (7.4)$$

showing that $n_p(s)$ divides $h(s)$. Similarly,

$$\begin{aligned} [1 - h(s)] &= w(s)n_p(s)d_p(s) - u_p(s)n_p(s) + 1 \\ &= w(s)n_p(s)d_p(s) + v_p(s)d_p(s) \\ &= [w(s)n_p(s) + v_p(s)]d_p(s) \end{aligned} \quad (7.5)$$

showing that $d_p(s)$ divides $[1 - h(s)]$.

Conversely, since $p(s) \neq 0$, $n_p(s) \neq 0$ and hence the unique candidate for $w(s)$ satisfying 7.3 is

$$w(s) = \frac{[u_p(s)n_p(s) - h(s)]}{n_p(s)d_p(s)} \quad (7.6)$$

which is stable under the hypothesis of the theorem. Indeed, since $n_p(s)$ and $d_p(s)$ are coprime and both $n_p(s)$ and $d_p(s)$ are non-zero

$$\frac{1}{n_p(s)d_p(s)} = \frac{v_p(s)}{n_p(s)} + \frac{u_p(s)}{d_p(s)} \quad (7.7)$$

hence

$$\begin{aligned} w(s) &= \left[\frac{v_p(s)}{n_p(s)} + \frac{u_p(s)}{d_p(s)} \right] [u_p(s)n_p(s) - h(s)] \\ &= v_p(s)u_p(s) + \frac{v_p(s)h(s)}{n_p(s)} + \frac{u_p(s)}{d_p(s)} [1 - h(s) - v_p(s)d_p(s)] \\ &= \frac{v_p(s)h(s)}{n_p(s)} + \frac{u_p(s)[1 - h(s)]}{d_p(s)} \end{aligned} \quad (7.8)$$

which is stable since $n_p(s)$ divides $h(s)$ and $d_p(s)$ divides $[1 - h(s)]$.

Note that if $p(s)$ is miniphase both $n_p(s)$ and $d_p(s)$ are miniphase, in which case the divisibility conditions of the theorem hold for all stable $h(s)$.

25. Corollary. If $p(s)$ is miniphase every stable $h(s)$ can be realized as the input-output gain of the feedback system of Figure 1 with a stabilizing compensator.

Of course, one can give a similar argument for the realization of any of the eight feedback system gains of Corollary 6. Note, however, that the $w(s)$ which simultaneously stabilizes the system and achieves the prescribed gain is unique and hence no further design latitude exists once such a gain has been realized.

26. Example. For the plant

$$p(s) = \frac{\left[\frac{(s+1)}{(s^2-4)} \right]}{\left[\frac{(s+2)}{(s+2)} \right]} = \frac{\left[\frac{(s+1)}{(s+2)^2} \right]}{\left[\frac{(s-2)}{(s+2)} \right]} = \frac{n_p(s)}{d_p(s)} \quad (7.9)$$

$n_p(s)$ is miniphase and thus divides every stable $h(s)$. As such, the only constraint on $h_{r_{2\mu_1}}(s)$ for a feedback system built around this plant is that $(s-2)/(s+2)$ divides $h_{r_{2\mu_1}}(s)$. That is $h_{r_{2\mu_1}}(s)$ must be stable and have a zero at $s = 2$.

Of course, in applications $p(s)$ is only an approximation to the actual plant, $\bar{p}(s)$, and, as such, it behooves us to investigate the degree to which our design based on nominal plant information remains valid for the actual plant. To this end we use the nominal plant model to design the compensator using the $w(s)$ of Equation 7.8 and then substitute the resultant $n_c(s)$ and $d_c(s)$ into the formula for $\tilde{h}_{r_{2\mu_1}}(s)$ given in Equation 2.10 with the actual plant model defining $n_{\bar{p}}(s)$ and $d_{\bar{p}}(s)$. A little algebra will then yield

$$\tilde{h}_{r_{2\mu_1}}(s) = \frac{h(s)n_{\bar{p}}(s)}{d_{\bar{p}}(s)d_c(s) + n_{\bar{p}}(s)n_c(s)n_p(s)}$$

Since $[d_{\bar{p}}(s)d_c(s) + n_{\bar{p}}(s)n_c(s)]$ is 1 when $\bar{p}(s) = p(s)$ the expression of Equation 7.10 reduces to $\tilde{h}_{r_{2\mu_1}}(s) = h(s)$ when $\bar{p}(s) = p(s)$. Moreover, since $n_p(s)$ divides $h(s)$ and $[d_{\bar{p}}(s)d_c(s) + n_{\bar{p}}(s)n_c(s)]$ is miniphase when $\bar{p}(s)$ is close to $p(s)$, $\tilde{h}_{r_{2\mu_1}}(s)$ is stable when $\bar{p}(s)$, as is the feedback system itself. Finally, since the inverse of a miniphase function is continuous in a neighborhood of 1, $\tilde{h}_{r_{2\mu_1}}(s)$ is not independent of $\bar{p}(s)$ in a neighborhood of $p(s)$. Thus, even though $\tilde{h}_{r_{2\mu_1}}(s)$ is not independent of $\bar{p}(s)$ the system is well behaved for $\bar{p}(s)$ in a neighborhood of $p(s)$ and small changes in the plant will manifest themselves in small changes in $\tilde{h}_{r_{2\mu_1}}(s)$.

Finally, we observe that in the process of realizing a prescribed input-output gain exactly we have used all available design latitude. As an alternative one might choose to only partially specify $\tilde{h}_{r_{2\mu_1}}(s)$. Indeed, this can be done allowing one more freedom in the choice of $w(s)$ and simultaneously permitting the constraints on $n_p(s)$ and $d_p(s)$ to be relaxed [3]. Although such a design can be implemented by characterizing those $w(s)$ which achieve the prescribed partial specifications it is often more convenient to reformulate the problem in an appropriate abstract ring whose elements achieve the prescribed partial specification. In particular, the *pole placement problem* can be resolved by such an approach [8]. As such, we will put off our investigation of this problem to the concluding section, wherein it will be discussed in the context of a ring theoretic generalization of our theory.

8. Pole placement

Although the result of the previous section yields a definitive criterion for the exact realization of a prescribed transfer function while simultaneously stabilizing the system the resultant criterion may be too restrictive. Indeed, for all practical purposes the transfer function design theorem requires that $p(s)$ be miniphase¹. Moreover, in the case the compensator is completely determined by $h(s)$ leaving no further design latitude. As such, we prefer to work with a less restrictive *pole placement problem* wherein we desire to

find a stabilizing compensator such that

$$h_{r_2\mu_1}(s) = \frac{r(s)}{q(s)} \quad (8.1)$$

where $q(s)$ is a prescribed Hurwitz denominator polynomial and $r(s)$ is an arbitrary numerator polynomial which has no common factors with $q(s)$ and whose order is less than or equal to that of $q(s)$ (so as to guarantee that $h_{r_2\mu_1}(s)$ will have no poles at infinity).

Since $q(s)$ is Hurwitz (for stability) any common factors between $r(s)$ and $q(s)$ will be in the left half plane and, as such, they must be characterized by polynomial coprimeness criterion rather than our stable rational function criterion which only characterizes common factors in the right half-plane. As such, we will use the term *polynomial coprime* to distinguish this criterion for our usual stable rational function *coprimeness*. Consistent with our use of the polynomial coprimeness concept we assume that

$$n_p(s) = \frac{a^l(s)a^r(s)}{m(s)} \quad (8.2)$$

and

$$d_p(s) = \frac{b^l(s)b^r(s)}{m(s)} \quad (8.3)$$

where $m(s)$ is a Hurwitz polynomial characterizing the poles of $n_p(s)$ and $d_p(s)$; $a^l(s)$ and $b^l(s)$ are Hurwitz polynomials characterizing the (strict) left half-plane zeros of $n_p(s)$ and $d_p(s)$, respectively; while $a^r(s)$ and $b^r(s)$ are anti-Hurwitz polynomials characterizing the finite (closed) right half-plane zeros of $n_p(s)$ and $d_p(s)$, respectively. Since $n_p(s)$ and $d_p(s)$ have no poles at infinity

$$o(m) \geq o(a^l) + o(a^r) \quad (8.4)$$

and

$$o(m) \geq o(b^l) + o(b^r) \quad (8.5)$$

where $o(\)$ denotes the order of the appropriate polynomial. Moreover, since $n_p(s)$ and $d_p(s)$ are coprime they cannot have a common zero at $s = 0$ and, as such, one of the inequalities, 8.4 or 8.5, must be satisfied with equality. Indeed, if we let

$$k = |o(a^l) + o(a^r) - o(b^l) - o(b^r)| \quad (8.6)$$

then either

$$o(m) = o(a^l) + o(a^r) = o(b^l) + o(b^r) + k \quad (8.7)$$

or

$$o(m) = o(a') + o(a'') + k = o(b') + o(b'') \quad (8.8)$$

Using this notation we obtain the following fundamental result.

Property. Given $p(s)$ there exists a compensator for the feedback system of Figure 1 which stabilizes the system and yields $h_{r, \mu_1}(s) = r(s)/q(s)$ where $q(s)$ is a prescribed Hurwitz polynomial, $r(s)$ is an arbitrary polynomial such that $o(r) \leq o(q)$, and $q(s)$ and $r(s)$ are polynomial coprime if and only if the equation

$$\alpha(s)a'(s) + \beta(s)b'(s) = q(s)$$

 admits polynomial solutions, $\alpha(s)$ and $\beta(s)$, such that

- (i) $o(\alpha) \leq o(q) + o(a') - o(m)$
- (ii) $o(\beta) \leq o(q) + o(b') - o(m)$
- (iii) $\alpha(s)$ and $\beta(s)$ are polynomial coprime.

Proof. If there exist polynomials $\alpha(s)$ and $\beta(s)$ satisfying

$$\alpha(s)a'(s) + \beta(s)b'(s) = q(s) \quad (8.9)$$

 and (i)-(iii) we define $w(s)$ by

$$w(s) = \frac{\beta(s)m(s)}{q(s)b'(s)}u_p(s) - \frac{\alpha(s)m(s)}{q(s)a'(s)}v_p(s) \quad (8.10)$$

Now, $q(s)$ and $b'(s)$ are both Hurwitz polynomials while $o(\beta m) \leq o(qb')$ via (ii) verifying that $\beta(s)m(s)/q(s)b'(s)$ is a stable rational function. Similarly, the fact that $q(s)$ and $a'(s)$ are both Hurwitz together with (i) implies that $\alpha(s)m(s)/q(s)a'(s)$ is stable. As such, $w(s)$ is represented as the sum or products of stable functions and is thus stable.

Using the $w(s)$ of 8.10 to define a compensator for our system via the stabilization theorem we obtain

$$\begin{aligned} h_{r, \mu_1}(s) &= -w(s)n_p(s)d_p(s) + u_p(s)n_p(s) = \frac{r(s)}{q(s)} \quad (8.13) \\ &= \left[-\frac{\beta(s)m(s)}{q(s)b'(s)}u_p(s) - \frac{\alpha(s)m(s)}{q(s)a'(s)}v_p(s) \right] n_p(s)d_p(s) + u_p(s)n_p(s) \\ &= -\frac{\beta(s)m(s)n_p(s)d_p(s)u_p(s)}{q(s)b'(s)} + \frac{\alpha(s)m(s)n_p(s)d_p(s)v_p(s)}{q(s)a'(s)} + u_p(s)n_p(s) \\ &= -\frac{\beta(s)b'(s)n_p(s)u_p(s)}{q(s)} + \frac{\alpha(s)a'(s)d_p(s)v_p(s)}{q(s)} + u_p(s)n_p(s) \\ &= [q(s) - \beta(s)b'(s)] \frac{n_p(s)u_p(s)}{q(s)} + \frac{\alpha(s)a'(s)d_p(s)v_p(s)}{q(s)} \end{aligned}$$

$$\begin{aligned}
 &= \frac{\alpha(s)a'(s)}{q(s)}n_p(s)u_p(s) + \frac{\alpha(s)a'(s)}{q(s)}d_p(s)v_p(s) \\
 &= \frac{\alpha(s)a'(s)}{q(s)}[u_p(s)n_p(s) + v_p(s)d_p(s)] \\
 &= \frac{\alpha(s)a'(s)}{q(s)} \tag{8.11}
 \end{aligned}$$

To verify that this is the required transfer function we must show that $r(s) = \alpha(s)a'(s)$ and $q(s)$ are polynomial coprime and that $o(r) \leq o(q)$. Now, it follows from 8.4 and (i) that

$$\begin{aligned}
 o(r) &= o(\alpha) + o(a') \leq o(\alpha) + o(m) - o(a^l) \\
 &\leq o(q) + o(a^l) - o(m) + o(m) - o(a^l) = o(q) \tag{8.12}
 \end{aligned}$$

verifying that $h_{r,2u_1}(s)$ has no poles at infinity. On the other hand if $r(s) = \alpha(s)a'(s)$ and $q(s)$ are not polynomial coprime, say they have a common zero at $s = s_0$, then 8.9 implies that $\beta(s)b'(s)$ also has a zero at s_0 . Moreover, since $q(s)$ is Hurwitz, s_0 must lie in the strict left half-plane implying that it is actually a common zero of $\alpha(s)$ and $\beta(s)$ since the zeros of $a'(s)$ and $b'(s)$ are in the right half-plane. This is, however, impossible since $\alpha(s)$ and $\beta(s)$ are assumed to be polynomial coprime. As such, $r(s)$ and $q(s)$ are polynomial coprime completing our sufficiency proof.

To verify the necessity of conditions (i)-(iii) assume that there exists a stable $w(s)$ such that

$$h_{r,2u_1}(s) = -w(s)n_p(s)d_p(s) + u_p(s)n_p(s) = \frac{r(s)}{q(s)} \tag{8.13}$$

Now, substitution of 8.2 and 8.3 into 8.13 implies that

$$u_p(s) = w(s)d_p(s) + \frac{r(s)m(s)}{q(s)a'(s)a^l(s)} \tag{8.14}$$

but since $u_p(s)$ is stable while $m(s)$, $q(s)$ and $a^l(s)$ are Hurwitz the anti-Hurwitz polynomial $a'(s)$, in 8.14 must be cancelled by a corresponding factor in $r(s)$. As such 8.14 implies that there exists a polynomial $\alpha(s)$ such that

$$r(s) = \alpha(s)a'(s) \tag{8.15}$$

Moreover, since $u_p(s)$ is stable and given by $u_p(s) = r(s)m(s)/q(s)a^l(s)$

$$o(\alpha) \leq o(q) + o(a^l) - o(m) \tag{8.16}$$

verifying (i). Similarly; it follows from the equality $u_p(s)n_p(s) + v_p(s)d_p(s) = 1$ together with 8.13 and 8.15 that

$$v_p(s) + w(s)n_p(s) = \frac{[q(s) - \alpha(s)a'(s)]m(s)}{q(s)b'(s)b'(s)} \quad (8.17)$$

As before, since $v_p(s) + w(s)n_p(s)$ is stable while $q(s)$, $b'(s)$ and $m(s)$ are Hurwitz it follows that there exists a polynomial $\beta(s)$ such that

$$[q(s) - \alpha(s)a'(s)] = \beta(s)b'(s) \quad (8.18)$$

which is the desired equality. Furthermore, since

$$v_p(s) + w(s)n_p(s) = \frac{\beta(s)m(s)}{q(s)b'(s)} \quad (8.19)$$

is stable it follows that

$$o(\beta) \leq o(q) + o(b') - o(m) \quad (8.20)$$

verifying (ii). Finally, if s_0 were a common zero of $\alpha(s)$ and $\beta(s)$ it would also be a zero of $r(s) = \alpha(s)a'(s)$ and $q(s) = \alpha(s)a'(s) + \beta(s)b'(s)$ which is impossible since $r(s)$ and $q(s)$ are assumed to be polynomial coprime. As such, $\alpha(s)$ and $\beta(s)$ are polynomial coprime verifying (iii) and completing the proof.

Consistent with Property 27 the solution of our pole placement problem reduces to the solution of the linear polynomial equation

$$\alpha(s)a'(s) + \beta(s)b'(s) = q(s) \quad (8.21)$$

under constraints (i)-(iii) of Property 27. Now, since $n_p(s)$ and $d_p(s)$ are coprime they have no common right half-plane zeros hence $a'(s)$ and $b'(s)$ are polynomial coprime which implies that 8.21 is soluable. It is not, however, clear that 8.21 represents a solution satisfying constraints (i)-(iii) of Property 27. Indeed, 8.21 represents $o(q) + 1$ equations in $o(\alpha) + o(\beta) + 2$ unknowns hence for a solution we require that

$$o(q) + 1 \leq o(\alpha) + o(\beta) + 2 \quad (8.22)$$

while constraints (i) and (ii) together with 8.7 (or 8.8) imply that

$$\begin{aligned} o(q) + 1 &\leq [o(q) + o(a') - o(m) + o(q) + o(b') - o(m)] + 2 \\ &= 2o(q) - o(a') - o(b') - k + 2 \end{aligned} \quad (8.23)$$

or equivalently

$$o(q) \geq [o(a') + o(b') + k] - 1 \quad (8.24)$$

As such, Equation 8.24 yields a necessary and sufficient condition for the linear equation 8.21 to be generically solvable. Indeed, the polynomial coprimeness of $a'(s)$ and $b'(s)$ implies that 8.21 will always admit coprime solutions when 8.24 is satisfied (though the equation may also admit solutions which are not polynomial coprime). On the other hand if 8.24 is not satisfied 8.21 is generically unsolvable though it may admit solutions for certain $q(s)$ which lie in an appropriately lower dimensional subspace.

Finally, let us denote the integer $[o(a') + o(b') + k]$ by $\pi(p)$ and observe that it represents the total number of (closed) right half-plane poles and zeros of $p(s)$. Indeed, $o(a')$ is precisely the number of finite closed right half-plane zeros of $p(s)$ while $o(b')$ is the number of finite poles of $p(s)$ and k represents the number of poles or zeros at infinity (depending on whether 8.7 or 8.8 holds). As such, $\pi(p)$ is a natural measure of the *degree to which $p(s)$ fails to be miniphase*. Consistent with the above we have verified the following.

Pole Placement Theorem: Given $p(s)$ there exists a compensator for the feedback system of Figure 1 which stabilizes the system and yields

$$h_{r_2\mu_1}(s) = \frac{r(s)}{q(s)}$$

where $q(s)$ is a prescribed Hurwitz polynomial, $r(s)$ is an arbitrary polynomial such that $o(r) \leq o(q)$, and $q(s)$ and $r(s)$ are polynomial coprime if

$$o(q) \geq \pi(p) - 1$$

Conversely, if $o(q) < \pi(p) - 1$ the pole placement problem is generically unsolvable.

Finally, we consider the problem of parameterizing the solution space for the pole placement problem. Indeed, it follows from Property 27 that the space of compensators which resolve the pole placement problem for a given $q(s)$ take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]} \quad (8.25)$$

where $w(s)$ is given by 8.10 with $\alpha(s)$ and $\beta(s)$ chosen to satisfy 8.21 under constraints (i)-(iii) of Property 27 together with the condition that $w(s)n_p(s) + v_p(s)$ is not identically zero. Now, if $\alpha^p(s)$ and $\beta^p(s)$ represent a particular solution to 8.21 it follows from the polynomial coprimeness of $a'(s)$ and $b'(s)$ together with the usual arguments that the entire solution space for 8.21 takes the form

$$\alpha(s) = \alpha^p(s) + j(s)b'(s) \quad (8.26)$$

and

$$\beta(s) = \beta^p(s) - j(s)a'(s) \quad (8.27)$$

where $j(s)$ is an arbitrary polynomial. In our case, however, in order to satisfy constraints (i) and (ii) of Property 27 we assume that $\alpha^p(s)$ and $\beta^p(s)$ satisfy these constraints in which case we may guarantee that $\alpha(s)$ and $\beta(s)$ will also satisfy constraints (i) and (ii) by requiring that

$$o(j) \leq o(q) - \pi(p) \quad (8.28)$$

(which fact follows from Equation 8.7 and 8.8)⁴. In general, however, these solutions may or may not be polynomial coprime. As such, we may parameterize the solution space for the pole placement problem via:

28. Corollary. Let $p(s)$ and $q(s)$ be given and let $\alpha^p(s)$ and $\beta^p(s)$ be solutions of Equation 8.21 such that $o(\alpha) \leq o(q) + o(a') - v(m)$ and $o(\beta) \leq o(q) + o(b') - o(m)$. Then the set of compensators which resolve the pole placement problem take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where

$$w(s) = \frac{[\beta^p(s) - j(s)a'(s)]m(s)}{[q(s)b'(s)]} u_p(s) - \frac{[\alpha^p(s) + j(s)b'(s)]m(s)}{[q(s)a'(s)]} v_p(s)$$

where $j(s)$ is an arbitrary polynomial such that $o(j) \leq o(q) - \pi(p)$ and $j(s)$ is chosen so that $\alpha(s) = \alpha^p(s) + j(s)b'(s)$ and $\beta(s) = \beta^p(s) - j(s)a'(s)$ are polynomial coprime and $w(s)n_p(s) + v_p(s)$ is not identically zero.

29. Example. Once again consider our usual example with plant

$$p(s) = \frac{(s+1)}{(s^2-4)} \quad (8.29)$$

though in this case since the formulation for the pole placement problem requires that $n_p(s)$ and $d_p(s)$ have a common denominator, $m(s)$, we will work with the fractional representation

$$p(s) = \frac{\left[\frac{(s+1)}{(s+3)^2} \right]}{\left[\frac{(s^2-4)}{(s+3)^2} \right]} = \frac{n_p(s)}{d_p(s)} \quad (8.30)$$

which is coprime since

$$\left[\frac{(6+13)}{(s+1)} \right] \left[\frac{(s+1)}{(s+3)^2} \right] + \left[1 \right] \left[\frac{(s^2-4)}{(s+3)^2} \right] = u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (8.31)$$

As such, for this example $m(s) = (s+3)^2$, $a'(s) = (s+1)$, $a''(s) = 1$, $b'(s) = (s+2)$, and $b''(s) = (s-2)$ while $\pi(p) = 2$. According to the theorem we may thus realize any Hurwitz denomination of order greater than or equal to 1.

Now, consider the case where we let $q(s) = s^2 + s + 1$ in which case we require that

$$o(\alpha) \leq o(q) + o(a') - o(m) = 1 \quad (8.32)$$

and

$$o(\beta) \leq o(q) + o(b') - o(m) = 1 \quad (8.33)$$

As such, 8.21 is satisfied by

$$[(3s+1)][1] + [s][(s-2)] = \alpha(s)a'(s) + \beta(s)b'(s) = q(s) = s^2 + s + 1 \quad (8.34)$$

where $\alpha(s)$ and $\beta(s)$ are both first order and polynomial coprime. Now, if we take $j(s) = 0$ in Corollary 28 we obtain

$$w(s) = \left[\frac{(s+3)^2(3s^2+6s-2)}{(s^2+s+1)(s+1)(s+2)} \right] \quad (8.35)$$

$$c(s) = \left[\frac{3(s+2)(-s^3+2s^2+11s+3)}{s(s+1)(s+3)^2} \right] \quad (8.36)$$

and

$$h_{v_2\mu_1}(s) = \left[\frac{3(s+1)}{(s^2+s+1)} \right] = \frac{\alpha(s)a'(s)}{q(s)} \quad (8.37)$$

as required.

9. Simultaneous stabilization

Although a robust design will remain valid under small plant perturbations one often desires to design a compensator which will meet specifications over a wide range of plants. For instance, if the plant contains unknown or variable parameters; say corresponding to temperature, altitude or load changes; one might wish to design a compensator which meets specifications independently of the unknown parameters. Alternatively, it may be required to design a compensator for a multi-mode plant which achieves the prescribed design specifications in each mode; say, a two speed motor. This

problem which has received renewed attention in recent years is termed the *simultaneous design problem* [7,17]. In the case of the simultaneous stabilization problem a complete but highly abstract theory exists.

In the present section, rather than summarizing this abstract theory, we will restrict ourselves to the one special case for which a simple frequency domain theory exists—the simultaneous stabilization of two plants [7]. Here, we assume that two distinct plants, $\bar{p}(s) \neq \hat{p}(s)$, are given and we desire to find a single compensator, $c(s)$, which simultaneously stabilizes both plants. We therefore assume that both $\bar{p}(s)$ and $\hat{p}(s)$ are characterized by coprime fractional representations

$$\bar{p}(s) = \frac{n_{\bar{p}}(s)}{d_{\bar{p}}(s)} \quad (9.1)$$

and

$$\hat{p}(s) = \frac{n_{\hat{p}}(s)}{d_{\hat{p}}(s)} \quad (9.2)$$

where

$$u_{\bar{p}}(s)n_{\bar{p}}(s) + v_{\bar{p}}(s)d_{\bar{p}}(s) = 1 \quad (9.3)$$

and

$$u_{\hat{p}}(s)n_{\hat{p}}(s) + v_{\hat{p}}(s)d_{\hat{p}}(s) = 1 \quad (9.4)$$

for appropriate stable functions $u_{\bar{p}}(s)$, $v_{\bar{p}}(s)$, $u_{\hat{p}}(s)$, and $v_{\hat{p}}(s)$.

Now it follows from the stabilization theorem that the stabilizing compensators for the two plants are given by

$$\bar{c}(s) = \frac{[-\tilde{w}(s)d_{\bar{p}}(s) + u_{\bar{p}}(s)]}{[\tilde{w}(s)n_{\bar{p}}(s) + v_{\bar{p}}(s)]} \quad (9.5)$$

and

$$\hat{c}(s) = \frac{[-\hat{w}(s)d_{\hat{p}}(s) + u_{\hat{p}}(s)]}{[\hat{w}(s)n_{\hat{p}}(s) + v_{\hat{p}}(s)]} \quad (9.6)$$

where $\tilde{w}(s)$ and $\hat{w}(s)$ are stable functions such that $\tilde{w}(s)n_{\bar{p}}(s) + v_{\bar{p}}(s)$ and $\hat{w}(s)n_{\hat{p}}(s) + v_{\hat{p}}(s)$ are not identically zero. In our application, however, we desire to construct a single compensator, $c(s)$, for both plants and hence we must equate 9.5 and 9.6. Recalling that the above coprime fractional representations for the compensator are unique up to a miniphase factor (via Property 1) we equate the numerators and denominators of 9.5 and 9.6 modulo a miniphase factor, $m(s)$, obtaining

$$-\tilde{w}(s)d_{\bar{p}}(s) + u_{\bar{p}}(s) = m(s)[- \hat{w}(s)d_{\hat{p}}(s) + u_{\hat{p}}(s)] \quad (9.7)$$

and

$$\tilde{w}(s)n_{\tilde{p}}(s) + v_{\tilde{p}}(s) = m(s)[\tilde{w}(s)n_{\tilde{p}}(s) + v_{\tilde{p}}(s)] \quad (9.8)$$

which must be solved for stable $\tilde{w}(s)$ and $\hat{w}(s)$, and miniphase $m(s)$. Rearranging the above equations yields the matrix equality

$$\begin{bmatrix} -d_{\tilde{p}}(s) & d_{\tilde{p}}(s) \\ n_{\tilde{p}}(s) & n_{\tilde{p}}(s) \end{bmatrix} \begin{bmatrix} \tilde{w}(s) \\ m(s)\hat{w}(s) \end{bmatrix} = \begin{bmatrix} m(s)u_{\tilde{p}}(s) - u_{\tilde{p}}(s) \\ m(s)v_{\tilde{p}}(s) - v_{\tilde{p}}(s) \end{bmatrix} \quad (9.9)$$

Solving 8.9 via Cramers rule then yields

$$\begin{bmatrix} \tilde{w}(s) \\ m(s)\hat{w}(s) \end{bmatrix} = \frac{1}{[n_{\tilde{p}}(s)d_{\tilde{p}}(s) - n_{\tilde{p}}(s)d_{\tilde{p}}(s)]} \begin{bmatrix} m(s) - [n_{\tilde{p}}(s)u_{\tilde{p}}(s) + d_{\tilde{p}}(s)v_{\tilde{p}}(s)] \\ m(s)[n_{\tilde{p}}(s)u_{\tilde{p}}(s) + d_{\tilde{p}}(s)v_{\tilde{p}}(s)] - 1 \end{bmatrix} \quad (9.10)$$

As such, we must construct a miniphase $m(s)$ such that

$$\tilde{w}(s) = \frac{m(s) - [n_{\tilde{p}}(s)u_{\tilde{p}}(s) + d_{\tilde{p}}(s)v_{\tilde{p}}(s)]}{[n_{\tilde{p}}(s)d_{\tilde{p}}(s) - n_{\tilde{p}}(s)d_{\tilde{p}}(s)]} \quad (9.11)$$

and

$$\hat{w}(s) = \frac{[n_{\tilde{p}}(s)u_{\tilde{p}}(s) + d_{\tilde{p}}(s)v_{\tilde{p}}(s)] - \frac{1}{m(s)}}{[n_{\tilde{p}}(s)d_{\tilde{p}}(s) - n_{\tilde{p}}(s)d_{\tilde{p}}(s)]} \quad (9.12)$$

are stable.

Since both the numerator and denominator in the above expressions are stable $\tilde{w}(s)$ and $\hat{w}(s)$ will be stable if and only if the numerators in 9.11 and 9.12 are chosen to cancel any right half-plane zeros in $[n_{\tilde{p}}(s)d_{\tilde{p}}(s) - n_{\tilde{p}}(s)d_{\tilde{p}}(s)]$. That is, if

$$n_{\tilde{p}}(s_i)d_{\tilde{p}}(s_i) - n_{\tilde{p}}(s_i)d_{\tilde{p}}(s_i) = 0 \quad (9.13)$$

for some right half-plane s_i we must choose $m(s_i)$ such that

$$m(s_i) - [n_{\tilde{p}}(s_i)u_{\tilde{p}}(s_i) + d_{\tilde{p}}(s_i)v_{\tilde{p}}(s_i)] = 0 \quad (9.14)$$

and

$$[n_{\tilde{p}}(s_i)u_{\tilde{p}}(s_i) + d_{\tilde{p}}(s_i)v_{\tilde{p}}(s_i)] - \frac{1}{m(s_i)} = 0 \quad (9.15)$$

While these may, at first, appear to be contradictory requirements the validity of 9.13 and 9.14 automatically implies that of 9.15, and, as such, we may simply define $m(s_i)$ by 9.14 at those values of s_i where 9.13 holds. Although this is true with complete generality to simplify our arguments we

will assume that $d_{\tilde{p}}(s)$ and $d_{\tilde{p}}(s)$ are coprime; i.e., they have no common (closed) half-plane zeros.

30. Lemma. If $d_{\tilde{p}}(s)$ and $d_{\tilde{p}}(s)$ are coprime then 9.14 and 9.15 are uniquely satisfied by

$$m(s_i) = \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)}$$

whenever 9.13 holds for s_i in the (closed) right half-plane.

Proof. First let us show that under our coprimeness assumption both $d_{\tilde{p}}(s_i)$ and $d_{\tilde{p}}(s_i)$ are non-zero. If $d_{\tilde{p}}(s) = 0$ then 9.13 implies that $n_{\tilde{p}}(s_i)d_{\tilde{p}}(s_i) = 0$ which is impossible since both $n_{\tilde{p}}(s)$ and $d_{\tilde{p}}(s)$ are coprime with respect to $d_{\tilde{p}}(s)$. Thus, $d_{\tilde{p}}(s_i) \neq 0$ while a similar argument applies to $d_{\tilde{p}}(s_i)$. Now, substituting

$$m(s_i) = \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} \quad (9.16)$$

into 9.14 together with 9.13 yields

$$\begin{aligned} m(s_i) - [n_{\tilde{p}}(s_i)u_{\tilde{p}}(s_i) + d_{\tilde{p}}(s_i)v_{\tilde{p}}(s_i)] \\ &= \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} - \frac{n_{\tilde{p}}(s_i)d_{\tilde{p}}(s_i)u_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} + \frac{d_{\tilde{p}}(s_i)d_{\tilde{p}}(s_i)v_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} \\ &= \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} - \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} [u_{\tilde{p}}(s_i)n_{\tilde{p}}(s_i) + v_{\tilde{p}}(s_i)d_{\tilde{p}}(s_i)] \\ &= \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} - \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} = 0 \end{aligned} \quad (9.17)$$

via 9.3. Similarly,

$$\begin{aligned} [n_{\tilde{p}}(s_i)u_{\tilde{p}}(s_i) + d_{\tilde{p}}(s_i)v_{\tilde{p}}(s_i)] - \frac{1}{m(s_i)} \\ &= \frac{n_{\tilde{p}}(s_i)d_{\tilde{p}}(s_i)u_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} + \frac{d_{\tilde{p}}(s_i)d_{\tilde{p}}(s_i)v_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} - \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} \\ &= \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} [u_{\tilde{p}}(s_i)n_{\tilde{p}}(s_i) + v_{\tilde{p}}(s_i)d_{\tilde{p}}(s_i)] - \frac{d_{\tilde{p}}(s_i)}{d_{\tilde{p}}(s_i)} \end{aligned}$$

$$= \frac{d_{\bar{p}}(s_i)}{d_{\bar{p}}(s_i)} - \frac{d_{\bar{p}}(s_i)}{d_{\bar{p}}(s_i)} = 0 \quad (9.18)$$

via 8.4

Note that the lemma remains valid even without the coprimeness assumptions on $d_{\bar{p}}(s)$ and $d_p(s)$ though if both of these functions have a common right half-plane zero at s_i one must take $m(s_i) = n_p(s_i)u_p(s_i)$.

Consistent with the above and retaining our assumption that $d_{\bar{p}}(s)$ and $d_p(s)$ are coprime the resolution of our simultaneous stabilization problem reduces to the construction of a miniphase function $m(s)$, such that

$$m(s) = \frac{d_{\bar{p}}(s_i)}{d_p(s_i)} \quad (9.19)$$

at each closed right half-plane zero of $n_{\bar{p}}(s)d_p(s) - n_p(s)d_{\bar{p}}(s)$. Of course, since all of our rational functions have real coefficients the set of points (s_i, m_i) which we must interpolate is *symmetric* in the sense that (s_i, m_i) is in the set whenever (\bar{s}_i, \bar{m}_i) is in the set where the "overbar" notation indicates complex conjugation. The required *miniphase interpolation lemma* was originally derived by Youla [25] in the context of his study of the stable stabilization problem and is repeated here without proof.

Youla's Lemma: Let (s_i, m_i) be a finite symmetric set of complex 2-tuples. Then there exists a miniphase function with real coefficients which interpolates (s_i, m_i) if and only if the set of m_i corresponding to the s_i lying on the positive real axis all have the same sign.

Note that this condition is extremely weak since we need only check s_i which are on the positive real axis. Of course, since (s_i, m_i) is symmetric these m_i must be real though they need not all be of the same sign. The necessity of the condition is clear for if two such m_i were of different signs the continuity of $m(s)$ along the positive real axis would imply that $m(s)$ had a zero on the positive real axis and hence it would not be miniphase. The sufficiency of the condition which is far less obvious was verified by Youla et al. in Reference 9.

To apply Youla's lemma to our interpolation problem we must check the sign of $d_{\bar{p}}(s_i)/d_p(s_i)$ for those s_i which represent zeros of $n_{\bar{p}}(s)d_p(s) - n_p(s)d_{\bar{p}}(s)$ lying on the positive real axis (including the point at infinity). Since we are not interested in the value of $d_{\bar{p}}(s_i)/d_p(s_i)$ but only its sign the desired information can be obtained by looking at the zero crossings of the functions $d_{\bar{p}}(s)$ and $d_p(s)$ along the positive real axis. By hypothesis these functions have no common right half-plane zeros and, hence, as long as $d_{\bar{p}}(s)$ and $d_p(s)$ taken together have an even number of

zeros crossing between each positive real axis zero of $n_{\bar{p}}(s)d_{\bar{p}}(s) - n_{\hat{p}}(s)d_{\hat{p}}(s)$ the sign of $d_{\bar{p}}(s)/d_{\hat{p}}(s)$ will remain constant at these zeros. Finally, recognizing that the right half-plane zeros of $d_{\bar{p}}(s)$ and $d_{\hat{p}}(s)$ are just the right half-plane poles of $\bar{p}(s)$ and $\hat{p}(s)$ we obtain the following theorem [7].

Simultaneous Stabilization Theorem: Let $\bar{p}(s)$ and $\hat{p}(s)$ be distinct plants with coprime fractional representations $\bar{p}(s) = n_{\bar{p}}(s)/d_{\bar{p}}(s)$ and $\hat{p}(s) = n_{\hat{p}}(s)/d_{\hat{p}}(s)$ where $d_{\bar{p}}(s)$ and $n_{\bar{p}}(s)$ are coprime. Then there exists a compensator which simultaneously stabilizes both plants if and only if $\bar{p}(s)$ and $\hat{p}(s)$ taken together have an even number of poles between every pair of positive real axis zeros of $n_{\bar{p}}(s)d_{\hat{p}}(s) - n_{\hat{p}}(s)d_{\bar{p}}(s)$.

The conditions for simultaneous stabilization implied by the theorem are actually quite weak in that more often than not $n_{\bar{p}}(s)d_{\hat{p}}(s) - n_{\hat{p}}(s)d_{\bar{p}}(s)$ has less than two positive real axis zeros in which case the conditions of the theorem are trivially satisfied. Although our theorem and its proof is quite complex the resultant test for simultaneous stabilization is extremely simple as illustrated by the pole-zero plots of Figure 5. Here, the plot of Figure 5a corresponds to a pair of plants which admit simultaneous stabilization since there are an even number of poles between every pair of positive real axis zeros. On the other hand the plot of Figure 5b corresponds to a pair of plants which do not admit simultaneous stabilization since there is an odd number of poles between one pair of zeros. Note, in both of these examples the poles of $\bar{p}(s)$ and $\hat{p}(s)$ are treated together and we do not have to distinguish between the poles of one plant and those of the other.

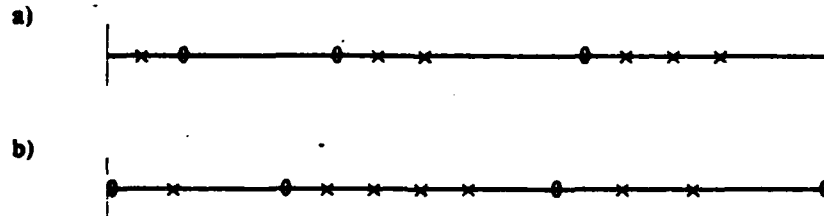


Figure 5. Pole-zero plots for the simultaneous stabilization problem.

Example. Once again consider the plant

$$\bar{p}(s) = \frac{(s+1)}{(s^2-4)} = \frac{\frac{(s+1)}{(s+2)^2}}{\frac{(s-2)}{(s+2)}} = \frac{n_{\bar{p}}(s)}{d_{\bar{p}}(s)} \quad (9.20)$$

and let us attempt to simultaneously stabilize this plant and

$$\bar{p}(s) = \left[\frac{(s+A)}{(s-1)} \right] = \frac{\left[\frac{(s+A)}{(s+2)} \right]}{\left[\frac{(s-1)}{(s+2)} \right]} = \frac{n_{\bar{p}}(s)}{d_{\bar{p}}(s)} \quad (9.21)$$

Now,

$$n_{\bar{p}}(s)d_{\bar{p}}(s) - n_{\bar{p}}(s)d_{\bar{p}}(s) = \frac{(A-2)s - (1+2A)}{(s+2)^2} \quad (9.22)$$

has zeros at

$$s_1 = \frac{(1+2A)}{(A-2)} \quad (9.23)$$

and $s_2 = \infty$ while the plants have poles at $s = 1$ and $s = 2$ as indicated in Figure 6. The conditions of the theorem are therefore satisfied if $s_1 > 2$ or $s_1 < 1$. Indeed, a little algebra will reveal that

$$s_1 > 2 \text{ if } A > 2 \quad (9.24)$$

while

$$s_1 < 1 \text{ if } -3 < A < 2 \quad (9.25)$$

which correspond to the two cases in which the two plants may be simultaneously stabilized. On the other hand

$$1 < s_1 < 2 \text{ if } A < -3 \quad (9.26)$$

To complete our example let us consider the case where $A = 3$ in which case $s_1 = 7$ and the two plants are simultaneously stabilizable. Indeed, to construct the required compensator we require a miniphase function, $m(s)$, such that

$$m(7) = \frac{d_{\bar{p}}(7)}{d_{\bar{p}}(7)} = \frac{6}{5} \quad (9.27)$$

and

$$m(\infty) = \frac{d_{\bar{p}}(\infty)}{d_{\bar{p}}(\infty)} = 1 \quad (9.28)$$

Since both values are positive Youla's lemma guarantees the existence of

such a function. Indeed,

$$m(s) = \frac{(s+5)}{(s+3)} \quad (9.29)$$

suffices. Finally upon substituting this choice of $m(s)$ into 9.11 the required $\tilde{w}(s)$ is obtained.

Unlike our previous design problems in which a complete parameterization of the appropriate family of compensators was obtained here we have only given an existence criteria and specified a method for constructing one such design when it exists. Although a family of such compensators exists no simple parameterization by the stable function is known. Indeed, it is apparent from the abstract solution to the general simultaneous stabilization problem of Reference 5 that no simple parameterization exists.

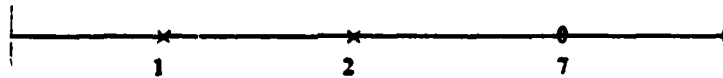


Figure 6. Positive real axis pole-zero plot for Example 31.

10. Stable stabilization

Occasionally, rather than simply designing a compensator to stabilize a feedback system we require that the compensator, itself, also be stable [25]. Unlike the general stabilization problem for which a solution always exists a *stable stabilizing compensator* may fail to exist. Fortunately, the problem of finding a stable stabilizing compensator is a special case of the *simultaneous stabilization problem* and, as such, we may derive criteria for the existence of a stable stabilizing compensator from the simultaneous stabilization theorem [7,25].

32. Lemma. A compensator $c(s)$ is stable if and only if it stabilizes the zero plant ($p(s) = 0$).

Proof. This follows immediately from Equations 2.2 and 2.4 upon observing that when $p(s) = 0$ the only non-trivial feedback system gain is $c(s)$.

Consistent with the lemma the problem of finding a stable stabilizing compensator for $p(s)$ is equivalent to the simultaneous stabilization problem with $\beta(s) = p(s)$ and $\hat{\beta}(s) = 0$. Thus upon letting $n_{\tilde{p}}(s) = n_p(s)$, $d_{\tilde{p}}(s) = d_p(s)$, $n_{\tilde{\beta}}(s) = 0$ and $d_{\tilde{\beta}}(s) = 1$ in the simultaneous stabilization theorem we obtain the following [7,25].

Stable Stabilization Theorem: Given $p(s)$ there exists a stable stabiliz-

ing compensator for the feedback system of Figure 1 if and only if an even number of poles of $p(s)$ lie between every pair of positive real axis zeros of $p(s)$.

Of course, one can construct a stable stabilizing compensator when it exists via the same technique used in the previous section to construct a solution to the simultaneous stabilization problem. That is, a miniphase function which interpolates $1/d_p(s)$ at the (closed) right half-plane zeros of $p(s)$ is constructed and used in Equation 9.11 to construct the appropriate $w(s)$. Also as in the case of the simultaneous stabilization problem no parameterization of the stable stabilizing compensators is known.

33. Example. Once again let us consider the plant

$$p(s) = \left[\frac{(s+1)}{(s^2-4)} \right] = \frac{\left[\frac{(s+1)}{(s+2)^2} \right]}{\left[\frac{(s-2)}{(s+2)} \right]} = \frac{n_p(s)}{d_p(s)} \quad (10.1)$$

where

$$\frac{16}{3} \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right] \left[\frac{(s-2)}{(s+2)} \right] = u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (10.2)$$

Now, the only positive real axis zero of $p(s)$ is at $s = \infty$ hence the existence of a stable stabilizing compensator is assured. To construct the required compensator we must find a miniphase function, $m(s)$, such that $m(\infty) = 1/d_p(\infty) = 1$. Clearly

$$m(s) = 1 \quad (10.3)$$

will suffice. Next we compute $w(s)$ via equation 8.11 which reduces to

$$w(s) = \frac{m(s) - v_p(s)}{n_p(s)} \quad (10.4)$$

for the stable stabilization problem. Substituting $m(s)$, $n_p(s)$ and $v_p(s)$ from 10.2 and 10.3 into 10.4 now yields

$$w(s) = \frac{4(s+2)}{3(s+1)} \quad (10.5)$$

Finally, this value of $w(s)$ is substituted into the compensator formula for this plant given in Equation 2.37. After some algebra we conclude that

$$c(s) = \frac{(20s+8)}{3(s+1)} \quad (10.6)$$

is the required stable stabilizing compensator.

11. Optimization of system performance

The final step in the feedback system design process is the use of any design latitude which remains after all design constraints have been achieved to optimize the system performance. This may represent a classical optimal control type of minimization using a weighted sum of the regulation error and the norm of the plant input or it may represent a more qualitative measure of system performance; reliability, sensitivity, etc. Needless to say the precise parameter(s) which one might desire to optimize is highly dependent on the physical system under design and, thus, beyond the realm of the present discussion.

Whatever parameter one chooses to optimize, the key to the optimization process lies with the simple affine nature of the feedback system gains formulated in Corollary 6. Indeed, each of these gains is affine in the design parameter $w(s)$. Moreover, when one includes a tracking or disturbance rejection constraint the subset of allowed $w(s)$ is also affine in the new design parameter, say $g(s)$. Since the composition of affine functions is affine the system gains remain affine in the stable design parameter, i.e.

$$h(s) = x(s)g(s) + y(s) \quad (11.1)$$

where $h(s)$ is the appropriate gain, $x(s)$ and $y(s)$ are given stable functions, and $g(s)$ is a stable design parameter. As such, the optimization of one or more of the system gains and/or the system responses to a specified class of inputs usually proves to be a straightforward process.

In particular, if one assumes that μ_1 and μ_2 are specified stochastic processes and desires to minimize the expected value of a weighted sum of system responses a classical Wiener-Hopf optimization problem results [8,23,24]. Similarly, the various deterministic optimization problems one might choose to investigate are greatly simplified by the affine nature of 10.1 which permits the norm of any system response to be expressed as a quadratic function of $x(s)$. As such, these problems are usually amenable to a simple completion of the squares solution.

Rather than going through a formal optimization process at this stage the system designer may simply choose to "optimize" some qualitative characteristic of the system; say, its *sensitivity*. Recall that the classical *Bode sensitivity measure* for a feedback system is just

$$\begin{aligned} S(s) &= \frac{1}{1 + p(s)c(s)} = h_{\mu_1}(s) \\ &= w(s)n_p(s)d_p(s) - v_p(s)d_p(s) \end{aligned} \quad (11.2)$$

hence to minimize the sensitivity of our feedback system it suffices to make $S(s)$ small in some sense. To this end one may go through a formal optimization procedure; say, a Chebychev minimization of $S(s)$ over a

prescribed frequency range; or simply choose $w(s)$ so that S has one or more imaginary axis zeros in the prescribed frequency range. Indeed, this is the case in the following example.

34. Example. For the plant

$$p(s) = \left[\frac{(s+1)}{(s^2-4)} \right] = \frac{\left[\frac{(s+1)}{(s+2)^2} \right]}{\left[\frac{(s-2)}{(s+2)} \right]} = \frac{n_p(s)}{d_p(s)} \quad (11.3)$$

where

$$\frac{16}{3} \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right] \left[\frac{(s-2)}{(s+2)} \right] = u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (11.4)$$

$S(s)$ takes the form

$$\begin{aligned} S(s) &= w(s) \left[\frac{(s+1)}{(s+2)^2} \right] \left[\frac{(s-2)}{(s+2)} \right] + \left[\frac{(s+2/3)}{(s+2)} \right] \left[\frac{(s-2)}{(s+2)} \right] \\ &= \frac{[w(s)(s+1) + (s^2 + 8/3s + 4/3)](s-2)}{(s+2)^3} \end{aligned} \quad (11.5)$$

Now, to minimize the sensitivity of our system we will choose a stable $w(s)$ which will place zeros on the imaginary axis in the frequency range of interest. If we take

$$w(s) = \frac{(As+B)}{(s+1)} \quad (11.6)$$

it will be stable for all choices of A and B and yield

$$S(s) = \frac{[s^2 + (A + 8/3)s + (B + 4/3)](s-2)}{(s+2)^3} \quad (11.7)$$

which will allow us to place a pair of zeros anywhere on the complex plane. For instance, if we let $A = -8/3$ and $B = -1/3$ 10.7 becomes

$$S(s) = \frac{[s^2 + 1](s-2)}{(s+2)^3} \quad (11.8)$$

which has zeros at $\pm j$. Of course, by using a more complex $w(s)$ additional zeros can be created.

12. Historical notes, generalizations, and conclusions

Although many of the concepts described in the preceding have a long

history in single-variate control theory, our purpose has been to use the single-variate case as a forum in which to survey the algebraic approach to the feedback system design problem developed during the past few years [4,8,11,16,17,23,24]. Indeed, a review of the preceding eleven sections will reveal that, with few exceptions, our theory has been formulated with no more complex mathematics than addition, subtraction, multiplication, and division. As such, one should not be surprised to find that most of the theory can be extended to the multivariate case [11,16,24] while much of it extends to the case of a general linear system; be it time-varying, distributed or multidimensional [4,8,10]. The purpose of this concluding section is therefore to review the results and literature pertinent to our algebraic approach to the feedback system design problem.

Although the polynomial fractional representation concept is implicit in any single-variate system theory it was not adapted to the multivariate case until the mid-sixties with the work of Rosenbrock, et al [2,15]. Since such a polynomial fractional representation does not admit an obvious generalization to the distributed case, several researchers began to search for an alternative during the mid-seventies, eventually settling on a fractional representation theory, wherein, an arbitrary system is represented as the ratio of two stable systems [4,8]. This was first used in the *distributed system theory* of Callier and Desoer [4], then extended to an *operator theoretic framework* [19], and finally to an abstract *ring theoretic setting* [8] in the late seventies.

The starting point for our theory, however, rests with the celebrated 1976 papers of Youla, Bongiorno, and Jabr [23,24] in which a complete parameterization of the set of stabilizing compensators for a general multivariate system was first formulated. Of course, such parameterizations have long been known in the single-variate case while several authors gave alternative formulations [1,3,5,6,9,13,14,15,20,21,26]. Although the *YBJ theory* was originally formulated using a polynomial fractional representation [23,24], it was soon discovered that the theory could be both simplified and generalized by working with stable factors rather than polynomial factors. This, in turn, led to both distributed [4] and ring theoretic [8] formulations of the stabilization theorem. More recently a *global formulation of the YBJ theory* has been given an *algebraic-geometric setting* [17].

Although the obvious motivation for formulating the feedback system design problem in a ring theoretic setting is to permit its generalization to *multivariate, time-varying, and distributed systems*, two additional benefits have, in fact, accrued from the theory. First, the ring theoretic setting forces one to adopt purely algebraic arguments which are often simpler than the analytic arguments used in the polynomial and rational function approaches. Second, rather than working with the full ring of stable systems one can work in a subring, thereby obtaining a design theory for "strongly

stable" systems [8]. In particular, this allows the YBJ stabilization theory to be immediately translated into a complete parameterization of the set of compensators which place the poles of a feedback system in a prescribed subset of the (strict) left half-plane. Indeed, the set of transfer functions with poles in such a subset form a ring which can be used to derive an alternative *pole placement theory* simply by representing $p(s)$ as the ratio of two functions with poles in the prescribed subset and similarly for $u_p(s)$, $v_p(s)$, etc. Unlike the formulation of Section 8, however, one cannot precisely specify the pole locations and multiplicities via such an approach.

In addition to the formal stabilization theory the YBJ theory also has had a significant impact on the "*philosophy of system design*". Indeed, with the advent of the YBJ theory the importance of parameterizing the entire family of systems which achieve the design constraints rather than simply specifying a single design was made apparent. In the original papers of Youla, Bongiorno, and Jabr the parameterization became the first step in a Wiener-Hopf optimization [23,24] over the set of stabilizing compensators. As such, the parameterization led to an explicit description of the constraint set for the optimization problem, thereby permitting one to carry out an unconstrained optimization over $w(s)$ rather than a constrained optimization over $c(s)$. Of course, the YBJ philosophy has been continued with the parameterization of the set of compensators which solve the various tracking and disturbance rejection problems [11,16]. Moreover, it has since been carried over to other fields of endeavor by Youla, who has parameterized the complete set of stochastic processes which are compatible with observed data [22]; and by Helton, who has given a complete parameterization of the set of *impedances* which are *compatible* with a prescribed load [12]. This, in turn, was then employed together with a *non-Euclidian optimization algorithm* to resolve a long standing *broadband matching problem*.

The *tracking and disturbance rejection problems* have their origins in classical single-variable control theory where the *final value theorem* is used to formulate the required criteria. The origin of the present formulation, however, rests with the work of Francis [10] who formulated a *divisibility criterion* for the solution of the tracking problem and the work of Callier and Desoer [4] who first integrated the YBJ philosophy with the tracking and disturbance rejection problem. The present formulation is, however, based on that of Sacks and Murray [16] who derived the design equations presented here in a general ring theoretic setting and formulated the coprimeness criteria for the existence of a solution in a multivariable setting.

Although *robustness* has been a topic of wide interest throughout control theory over the past decade the present formulation dates only to the work of Francis [10] in 1977 and is based on the more recent work of Francis and Vidyasagar [11]. To our knowledge, however, the explicit parameterization

of the compensators which resolve the *robust tracking and/or disturbance rejections problems* has not previously appeared.

Unlike the topics considered in Sections 1 through 8 which have been of interest to control theorists for a number of years the simultaneous design problem has only been recently formulated in an effort to develop an "*ultra robust*" control theory [7,17,19]. Indeed, the solution of the two plant problem presented here was first derived by Chua [7] and has not previously appeared in the literature while the two plant multivariate problem has been resolved by Vidyasagar and Viswanadham [19]. At the time of this writing the two plant problem is the only simultaneous stabilization problem for which an explicit frequency domain solution is known. Indeed, the techniques used herein do not extend, even to the three plant case, in any obvious manner [7]. Interestingly, however, a complete solution to the *general simultaneous stabilization problem* has been given in an abstract *algebraic-geometric setting* [17] though no practical method for implementing this theory is available. In essence, the simultaneous stabilization problem is a *global problem* in which one must characterize the properties of the entire set of plants one desires to stabilize rather than simply give a local parameterization of the elements of that set. As such, it has been necessary to formulate the problem in an abstract setting wherein the easy implementability of our theory is lost.

A solution to the stable stabilization problem was first presented by Youla, Bongiorno, and Lu [25], even before the publication of the YBJ theory, though it was essentially unknown until the relationship between the *stabilization problem* and the *simultaneous stabilization problem* was uncovered. Unlike the remainder of our theory the solution to the two plant simultaneous stabilization problem and the related solution of the stable stabilization problem are highly pole-zero oriented and, as such, it is unclear whether or not they admit a viable extension beyond the multivariable case. Interestingly, however, when the pole-zero interlacing property of Section 10 fails, one can determine the minimum number of unstable poles required by any stabilizing compensator simply by counting the number of times the plant has an odd number of poles between a pair of its positive real axis zeros [19]. Finally, we note that the "optimization" techniques are matched to our theory. Indeed, the relationship between the stabilization theory and the Wiener-Hopf theory was invoked in the original YBJ papers [23,24].

¹ This journal, CSSP, Vol. 1, No. 2, 1982, pages 137 to 169.

² To be rigorous one must invoke some sophisticated mathematics to prove that there exists a $d_p(s)$ which is coprime with $d_r(s)$ in every neighborhood of $d_p(s)$ [11]. As such, cancellation cannot be guaranteed for every $d_p(s)$ though some $d_p(s)$ may cancel the instabilities in $d_r(s)$.

³ Although the theorem allows for the possibility that right half-plane zeros of $p(s)$ are matched by corresponding zeros in $h(s)$ and right half-plane poles of $p(s)$ are matched by corresponding poles of $\{1 - h(s)\}$ such matching is not robust and thus, for all practical purposes the transfer function design theorem requires that $p(s)$ be miniphase.

⁴ If one assumes that $\sigma(q) \geq \pi(p) - 1$ as per the theorem then the bound $\sigma(q) - \pi(p)$ is greater than or equal to -1 . In particular, this bound is equal to -1 when $\sigma(q) = \pi(p) - 1$ in which case 8.21 admits a unique solution and no design freedom is allowed.

References

1. Antsaklis, P.J., and J.B. Pearson, "Stabilization and Regulation in Linear Multivariate Systems", IEEE Trans. on Auto. Cont., Vol. AC-23, pp. 928-930, (1978).
2. Barnett, S. *Matrices in Control Theory*, London, Van Nostrand, 1971.
3. Bengtsson, G., "Output Regulation and Internal Modes - A Frequency Domain Approach", Automatica, Vol. 13, pp. 333-345, (1977).
4. Callier, F.M., and C.A. Desoer, "Stabilization, Tracking and Disturbance Rejection in Linear Multivariable Distributed Systems" Proc. of the 17th IEEE Conf. on Decision and Control, San Diego, Jan. 1979, pp. 513-514, (also Tech Memo UCB&ERL M78&83, Univ. of California at Berkeley, Dec. 1978).
5. Chang, L., and J.B. Pearson, "Frequency Domain Synthesis of Multivariable Linear Regulators", IEEE Trans. on Auto. Cont., Vol. AC-23, pp. 3-15, (1978).
6. Cheng, L., and J.B. Pearson, "Synthesis of Linear Multivariable Regulators", IEEE Trans. on Auto. Cont., (to appear).
7. Chua, O., M.S. Thesis, Texas Tech Univ., 1980.
8. Desoer, C.A., Liu, R.-w., Murray, J., and R. Sacks, "Feedback System Design: the fractional representation approach to analysis and synthesis", IEEE Trans. on Auto. Cont., Vol. AC-25, pp. 399-412, (1980).
9. Feintuch, A., and R. Sacks, *System Theory: A Hilbert Space Approach*, New York, Academic Press, (to appear).
10. Francis, B., "The Multivariable Servomechanism Problem from the Input-Output Viewpoint", IEEE Trans. on Auto. Cont., Vol. AC-22, pp. 322-328, (1977).
11. Francis, B., and M. Vidyasagar, "Algebraic and Topological Aspects of the Servo Problem for Lumped Linear Systems", unpublished notes, Yale Univ., 1980.
12. Helton, J.W., "Orbit Structure of the Moebius Transformation Semigroup Acting on H " in *Topics in Functional Analysis*, Advances in Mat. Supp. Studies, Vol. 3, New York, Academic Press, 1978, pp. 129-157.
13. Pernebo, L., Ph.D. Thesis, Lund Inst. of Tech., 1978.
14. Pernebo, L., "An Algebraic Theory for Design of Controllers for Linear Multivariable Systems; Parts I and II", IEEE Trans. on Auto. Cont., (to appear).
15. Rosenbrock, H.H., *State-Space and Multivariable Theory*, New York, J. Wiley and Sons, 1970.
16. Sacks, R., and J. Murray, "Feedback Systems Design: the tracking and disturbance rejection problems", IEEE Trans. on Auto. Cont., Vol. AC-26, pp. 203-217, (1981).

17. Sacks, R., and J. Murray, "Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem", unpublished notes, Texas Tech Univ., 1980.
18. Vidyasagar, M., Schneider, H., and B. Francis, "Algebraic and Topological Aspects of Feedback System Stabilization", Tech Rpt. 80-90, Dept. of Elec. Engrg., Univ. of Waterloo, 1980.
19. Vidyasagar, M., and M. Viswanadham, "Algebraic Design Techniques for Reliable Stabilization", Tech Rpt. 81-02, Dept. of Elec. Engrg., Univ. of Waterloo, 1980.
20. Wolovitch, W.A., "Multivariable System Synthesis with Step Disturbance Rejection", IEEE Trans. on Auto. Cont., Vol. AC-19, pp. 127-130, (1974).
21. Wolovitch, W.A., and P. Ferreira, "Output Regulation and Tracking in Linear Multivariable Systems", IEEE Trans. on Auto. Cont., Vol. AC-24, pp. 460-465, (1979).
22. Youla, D.C., "Interpolary Multichannel Spectral Estimation", Unpublished Notes, Polytechnic Inst. of New York (1979).
23. Youla, D.C., Bongiorno, J.J., and H.A. Jabr, "Modern Wiener-Hopf Design of Optimal Controllers, Part I", IEEE Trans. on Auto. Cont., Vol. AC-21, pp. 3-15, (1976).
24. Youla, D.C., Bongiorno, J.J., and H.A. Jabr, "Modern Wiener-Hopf Design of Optimal Controllers, Part II", IEEE Trans. on Auto. Cont., Vol. AC-21, pp. 319-338, 1976.
25. Youla, D.C., Bongiorno, J.J., and C.N. Lu, "Single-Loop Feedback Stabilization of Linear Multivariable Dynamical Plants", Automatica, Vol. 10, pp. 159-173, (1974).
26. Zames, G., "Feedback and Optimal Sensitivity: Model reference transformation, weighted seminorms, and approximate inverses", IEEE Trans. on Auto. Cont., (to appear).

SIMULTANEOUS DESIGN OF CONTROL SYSTEMS

by

R. Saeks and J. Murray
Department of Electrical Engineering
Texas Tech University
Lubbock, Texas 79409

ABSTRACT

The problem of designing a feedback controller which stabilizes a number of plants simultaneously is discussed from the fractional representation point of view. An abstract solution of this general simultaneous stabilization problem is presented, and an elementary, explicit criterion is given for the simultaneous stabilizability of two systems. Finally, some examples and counter examples are presented, and some open problems are discussed.

1. INTRODUCTION

Classically, in control theory one is given a plant and desires to design a control system around this plant which meets certain design specifications. In fact, however, a "real world" plant is never known exactly and, as such, a realistic design must simultaneously meet specifications over an entire range of plants which (hopefully) include the actual plant. The simplest form of the resultant *simultaneous design problem* is the *robust design problem* wherein one desires to meet the design specifications in an ϵ -ball around a prescribed nominal plant. Although this is satisfactory for dealing with modeling errors it cannot cope with plants containing unknown parameters and/or plants characterized by multiple modes of operation. For instance, the dynamics of an airplane or rocket vary widely with altitude while the dynamics of an electric motor change with speed and load. To cope with these problems we must formulate a *simultaneous design theory* in which one designs a control system to simultaneously meet specifications over a prescribed set of plants. Of course, the set of plants may be taken to be a ball in which case the classical robustness theory is replicated. Alternatively, one may choose to work with a set of plants in which one or more parameters vary over a prescribed range and/or a discrete set of plants; say, the dynamics of a two speed motor in its high and low speed settings.

The purpose of this paper is to review the state of research on the simultaneous design problem including the derivation of an explicit criterion of the simultaneous stabilization of two distinct plants and an algebro-geometric solution of the general simultaneous stabilization problem. In addition, the fundamental relationship between the simultaneous stabilization problem and the problem of designing a stable (or minimally

unstable) stabilizing compensator for a given plant is reviewed.

2. SIMULTANEOUS STABILIZATION AND STABLE STABILIZATION

We consider the feedback system shown in Fig. 1; this is characterized by the connection equations

$$e_p = u_2 + v_c$$

$$e_c = u_1 - v_p$$

or

$$\begin{bmatrix} e_p \\ v_p \end{bmatrix} = Q \begin{bmatrix} e_c \\ v_c \end{bmatrix} + \begin{bmatrix} u_2 \\ u_1 \end{bmatrix}$$

where

$$Q = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad (2.1)$$

To describe the dynamics of the plant and controller, we use the abstract fractional representation theory of [1], [2], [3]. This assumes four sets

$$g \supset h \supset i \supset j$$

where g is a ring with identity which represents the general class of systems with which we wish to work, h is a subring of g corresponding to the stable systems in g , i is a multiplicative set consisting of the elements in h which have inverses in g , and j is the subgroup of h which consists of the elements of h which have inverses in h .

We assume that the plant P has a right-coprime fractional representation

$$P = N_r D_r^{-1}$$

where the coprimeness is exhibited by

$$U_r N_r + V_r D_r = 1.$$

and also a left-coprime representation

$$P = D_1 N_1^{-1}$$

with

$$N_1 U_1 + D_1 V_1 = 1.$$

It has been shown [2] that one can then find U_1 and V_1 so that, in addition,

$$V_1^* U_1 = U_1^* V_1.$$

Thus the plant P can be described by the matrix

$$R_p = \begin{bmatrix} D_r & -U_1 \\ N_r & V_1 \end{bmatrix}$$

or its inverse

$$S_p = \begin{bmatrix} V_r & U_r \\ -N_1 & D_1 \end{bmatrix}.$$

The admissible input-output pairs (c_p, v_p) for the plant are then described by

$$\begin{bmatrix} c_p \\ v_p \end{bmatrix} = R_p \begin{bmatrix} \sigma_p \\ 0 \end{bmatrix}$$

where c_p is a kind of "partial state" for the plant P .

The corresponding matrices for the controller will be denoted by R_c and S_c , etc..

It can then be shown [3], [4] that a given controller C will stabilize a plant P if and only if

$$R_p = Q R_c Q^T W \quad (2.2)$$

where Q is the connection matrix (2.1), and W is of the form

$$\begin{bmatrix} 1 & 0 \\ s & 1 \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} \\ 0 & e_{22} \end{bmatrix} \quad (2.3)$$

where each matrix is in $GL_2(h)$ - i.e., is a 2×2 matrix of elements of h which has an inverse whose elements are in h . It is easy to see that being of the form (2.3) is equivalent to being an element of $GL_2(h)$ whose $(1,1)$ -element is in j , and that this in turn is equivalent to being the R -matrix of a stable system. Thus in terms of the R -matrix representation, we can restate (2.2) as

Theorem 1: [4]

The set of plants stabilizable by a compensator C is precisely the set of stable plants transformed by left multiplication by $Q R_c Q^T$.

This can also be restated as follows: A set P of plants is simultaneously stabilizable iff P lies in an image of the stables under left multiplication by some element of $GL_h(2)$.

These criteria, while geometrically appealing, can not be handled analytically; the following equivalent criterion is therefore useful.

Theorem 2: [4]

Let P be a set of plants represented by R_p , $p \in P$. The set P is simultaneously stabilizable iff there exists a family of matrices W_p , $p \in P$, of the form (2.3) such that

$$W_p^{-1} W_q = R_p^{-1} R_q, \quad \forall p, q \in P.$$

Proof:

The necessity of this condition is obvious from (2.2). The sufficiency can be checked by defining

$$R_c^{-1} = Q^T R_p W_p^{-1} Q$$

for any $p \in P$, and using the condition to check that R_c is well defined. Then (2.2) follows.

One can insert the requirement that the compensator be stable merely by adjoining the zero plant to the given set of plants - a compensator is stable if and only if it stabilizes the zero plant. Thus the stable stabilization of n plants simultaneously can be treated as a problem of simultaneously stabilizing $n+1$ plants. The converse, although less obvious, is also true - see [3].

3. THE TWO-PLANT CASE

The only case of which we know in which even the analytic criterion in section 2 can be implemented is the case of two plants. In this case, the criterion is as follows: two plants, with representing matrices R_1 and R_2 , can be simultaneously stabilized iff there exist matrices W_1 and W_2 of the form (2.3) such that

$$W_1^{-1} W_2 = R_1^{-1} R_2 \quad (3.1)$$

We note that, since the identity R -matrix corresponds to the zero plant, this is equivalent to the condition that the "plant" $R_1^{-1} R_2$ have a stable stabilizing compensator. (This is an example of the converse mentioned at the end of section 2.) For the linear time-invariant case, the problem has been solved by Youla [5]. It is of interest, however, to relate Youla's solution

to our approach. To this end, we denote $R_1^{-1}R_2$ by

$$R = \begin{bmatrix} D & -U \\ N & V \end{bmatrix}$$

and restrict ourselves to the scalar-input scalar-output case. If we write (3.1) in the form

$$W_1 R = W_2$$

and use the fact that the (1,1)-entry in each of the W -matrices must be in j , we get the condition:

$$D + TN \in j \text{ for some } t \in h.$$

In the present case, this is clearly equivalent to requiring the existence of a stable, minimum-phase transfer function which, at each closed right half-plane zero of N , interpolates D to an order equal at least to the order of the zero of N . Continuity and realness of the transfer functions show that a necessary condition for this is that D have the same sign at all closed positive real-axis zeros of N . Youla showed that this was also sufficient.

Thus to solve the problem of simultaneously stabilizing two plants, it remains only to express N and D in terms of the original plants. An easy calculation shows that we can take

$$N = N_2 D_1 - N_1 D_2$$

and

$$D = V_1 D_2 + U_1 N_2$$

where

$$R_1 = \begin{bmatrix} D_1 & -U_1 \\ N_1 & V_1 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} D_2 & -U_2 \\ N_2 & V_2 \end{bmatrix}$$

Thus in order for the plants to be simultaneously stabilizable, it is necessary and sufficient that $V_1 D_2 + U_1 N_2$ have the same sign at all closed positive real-axis zeros of $N_2 D_1 - N_1 D_2$.

Some calculations involving the coprimeness conditions show that this is equivalent to the condition that D_2/D_1 have the same sign at all closed positive real-axis zeros of $N_2 D_1 - N_1 D_2$.

This gives a criterion in terms of the transfer-functions themselves [3][4].

4. EXAMPLES AND PROBLEMS

The only case in which the geometric results can be illustrated on paper is the case in which there are only two parameters. For this reason our first example will deal with this situation.

EXAMPLES: [4]

Suppose given a family of plants of the form

$$p(s) = \frac{A}{s+B}$$

we would like to know when it is possible to stabilize these simultaneously by use of a proportional compensator with gain t . In this case the denominator of the closed-loop transfer function is

$$d(s) = s + (B+tA)$$

and so the feedback system will be stable iff $B+tA > 0$. Thus a set of plants is simultaneously stabilizable iff for some t , $B+tA > 0$ for each plant in the set. Now each plant can be represented as a point in the (A,B) -plane; in this representation, a set of plants is simultaneously stabilizable iff the set lies entirely above some straight line through the origin. The slope of such a line is $-t$, where t is the gain of a stabilizing compensator. For example, the set in Figure 2.b is stabilizable, while the sets in Figures 2.c and 2.d are not. Since the set of stables in this case is the upper half-plane, this gives a very vivid (although very special) illustration of theorem 1: a set of plants is simultaneously stabilizable iff it is contained in a rotated (or more accurately, sheared) version of the set of stables.

Our second example is a counterexample to the effect that even if every pair of plants in a set is simultaneously stabilizable, the entire set may not be. To this end, we take a set consisting of three plants $\{p_0, p_1, p_2\}$; here p_0 is the zero plant, and

$$p_1 = n_1/d_1, \quad p_2 = n_2/d_2$$

where n_1, d_1, n_2, d_2 are graphed in Figure 3. It is easy to see, by using Youla's criterion, that p_1 and p_2 each have stable stabilizing compensators, and by using the criterion in section 3, that p_1 and p_2 have a simultaneous stabilizing compensator. However, there is no stable compensator which simultaneously stabilizes p_1 and p_2 . If there were, then by the criterion of Theorem 2 there would be a stable transfer function f such that both $d_1 + fn_1 = r_1$ and $d_2 + fn_2 = r_2$ were stable and minimum-phase. At the zeros of n_1 , $d_1 > 0$ and so r_1 must be positive on the positive real axis. Similarly r_2 must be positive on the positive real axis. However, if we eliminate f we see that

$$n_2 d_1 - n_1 d_2 = n_2 r_1 - n_1 r_2$$

and so, at the zeros of $n_2 d_1 - n_1 d_2$, we must have $n_2 r_1 = n_1 r_2$. But at these points, $n_1 > 0$ and $n_2 < 0$, and so we get a contradiction. Thus p_0, p_1

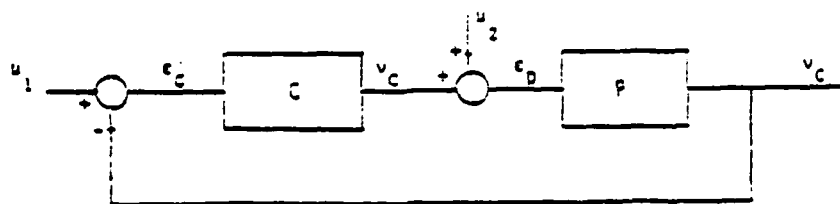


Figure 1. Feedback System

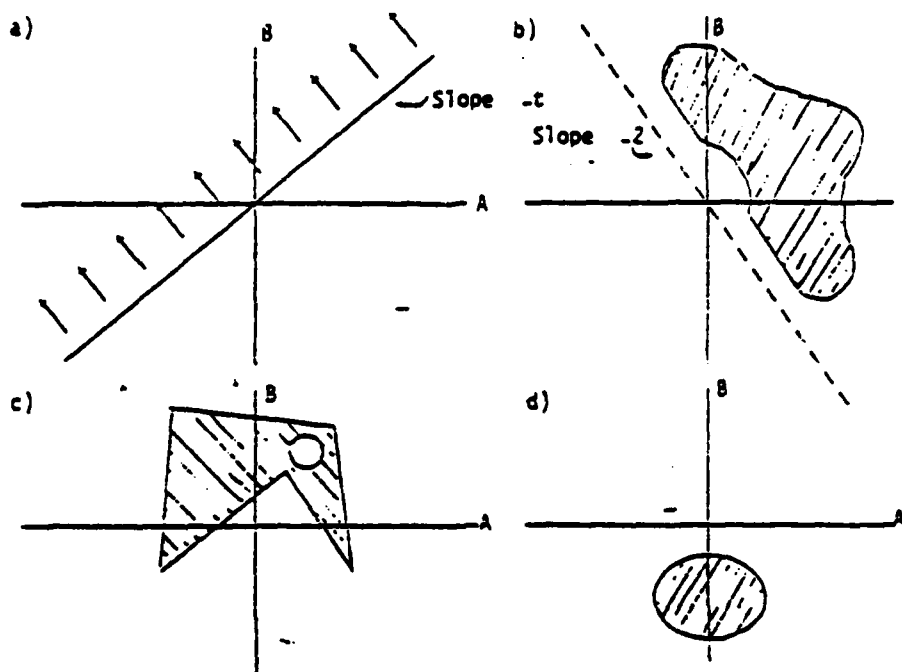


Figure 2. Simultaneous stabilization of a 1st order plant with a proportional compensator.

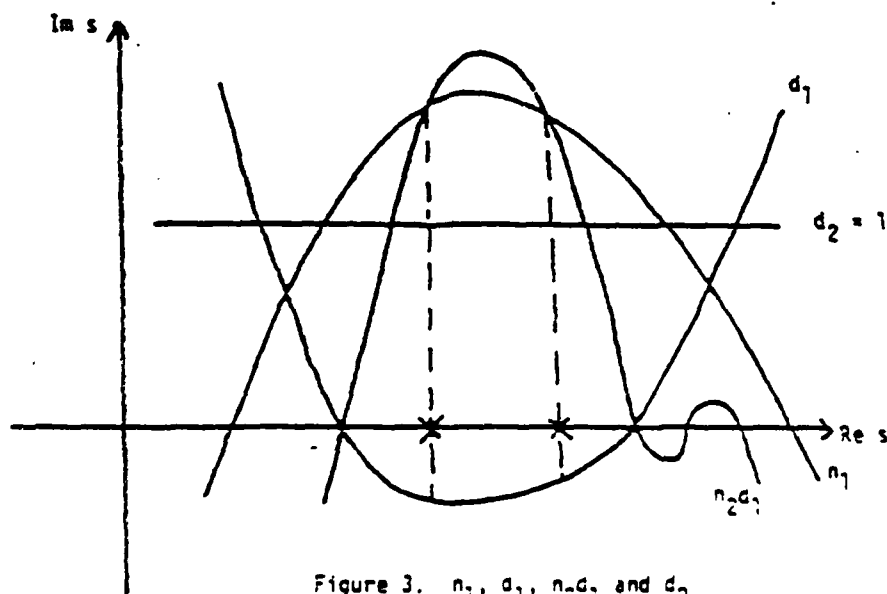


Figure 3. n_1 , d_1 , $n_2 d_1$ and d_2

and p_2 can not be simultaneously stabilizable.

It should be clear from the above that there are more problems than solutions in this area of research. At present, for example, we do not know any testable necessary and sufficient conditions for the simultaneous stabilizability of even three scalar plants. Thus as a first problem, we might state:

Problem 1: Find necessary and sufficient conditions for the simultaneous stabilizability of $N(>2)$ plants.

The following problem may also be of interest:

Problem 2: If one knows that a set of plants can be simultaneously stabilized, how does one find all compensators (or even some compensators) which will stabilize the set?

Problem 3: Since stabilization alone is usually not enough, can one find conditions for the existence of a compensator which, in addition to simultaneously stabilizing a set of plants, will also cause them to satisfy some other conditions (e.g. track a specified input signal)?

REFERENCES

1. Desoer, C.A., Liu, R.-w., Murray, J., and R. Saeks, "Feedback System Design: the Fractional Representation Approach to Analysis and Synthesis," IEEE Trans. Autom. Contr., AC-25 (1980) 399-412.
2. Saeks, R., and J. Murray, "Feedback System Design: the Tracking and Disturbance Rejection Problems," IEEE Trans. Autom. Contr., AC-26 (1981) 203-217.
3. Vidyasagar, M., and M. Viswanadham, "Algebraic Design Techniques for Reliable Stabilization," Report 81-02, Dept. of Elec. Eng., University of Waterloo.
4. Saeks, R., and Murray, J., "Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem," unpublished notes, Texas Tech Univ., 1980.
5. Youla, D.C., Bongiorno, J.J., and C.N. Lu, "Single-Loop Feedback Stabilization of Linear Multivariable Dynamical Plants," Automatica, 10 (1974), 159-173.

5. CONCLUSIONS

We have discussed the problem of finding a compensator which stabilizes every plant in a given set of plants. In the abstract, both geometric and analytic criteria have been given for the existence of such a compensator. However, only in very special cases, such as the case of two plants, can these criteria be checked. Another special case, in which the geometric criterion becomes particularly clear, is the case of first-order plants with proportional controllers. We have also given an example where pairwise simultaneous stabilizability of a set of plants does not imply overall simultaneous stabilizability. Finally, we have indicated some directions for further research.



Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem

RICHARD SAEKS, FELLOW, IEEE, AND JOHN MURRAY, MEMBER, IEEE

Abstract — An explicit relationship between the fractional representation approach feedback system design and the algebro-geometric approach to system theory is formulated and used to derive a global solution to the feedback system problem. These techniques are then applied to the simultaneous stabilization problem, yielding a natural geometric criterion for a set of plants to be simultaneously stabilized by a single compensator.

1. INTRODUCTION

CLASSICALLY, in control theory one is given a plant and desires to design a control system around this plant which meets certain design specifications. In fact, however, a "real world" plant is never known exactly and, as such, a realistic design must simultaneously meet specifications over an entire range of plants which (hopefully) include the actual plant. The simplest form of the resultant *simultaneous design problem* is the *robust design problem* wherein one desires to meet the design specifications in an ϵ ball around a prescribed nominal plant. Although this is satisfactory for dealing with modeling errors, it cannot cope with plants containing unknown parameters and/or plants characterized by multiple modes of operation. For instance, the dynamics of an airplane or rocket vary widely with altitude, while the dynamics of an electric motor change with speed and load. To cope with these problems, we must formulate a *simultaneous design theory* in which one designs a control system to simultaneously meet specifications over a prescribed set of plants. Of course, the set of plants may be taken to be a ball, in which case the classical robustness theory is replicated. Alternatively, one may choose to work with a set of plants in which one or more parameters vary over a prescribed range and/or a discrete set of plants, say, the dynamics of a two-speed motor in its high- and low-speed settings.

The simultaneous design concept is possibly best illustrated in the first-order case wherein a simple geometric solution suggests itself. Assume that our plants are of the form

$$p(s) = \frac{A}{s+B} \quad (1.1)$$

Manuscript received December 5, 1980; revised November 9, 1981. Paper recommended by E. W. Kamen, Chairman of the Linear Systems Committee. This work was supported in part by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

The authors are with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

and we desire to design a stable feedback system using a proportional compensator with gain t . This results in a system with characteristic function

$$d(s) = s + (B + tA) \quad (1.2)$$

and, as such, the feedback system will be stable if and only if $B + tA > 0$. Here, for a given compensator t , the feedback system will be stable if and only if the point (A, B) lies above the line with slope $1/t$ as shown in Fig. 1(a). As such, if we want to simultaneously stabilize an entire set of plants, their representations on the A - B plane must all lie above a line through the origin. For instance, the set of plants indicated by the hatched region in Fig. 1(b) can be simultaneously stabilized (by a compensator with gain $-1/2$), while the set of plants shown in Fig. 1(c) cannot be simultaneously stabilized since they subtend an angle greater than 180° on the A - B plane. Similarly, the set of plants shown in Fig. 1(d) cannot be simultaneously stabilized since they cross the negative A axis.

The above example suggest two alternative criteria for the simultaneous stabilization problem. One may adopt an algebraic criterion to the effect that

$$B + tA > 0 \quad (1.3)$$

for each plant in the prescribed set and some t . While such a test is definitive, it is local in nature, allowing one to test for stabilizability on a plant-by-plant basis, but yielding no global criterion with which to characterize a set of plants which is simultaneously stabilizable. To the contrary, one may adopt a global geometric viewpoint to the effect that a prescribed set of plants is simultaneously stabilizable if and only if it is contained in an appropriate half-plane. The goal of the present paper is the formulation of a similar geometric criterion for the simultaneous stabilization problem applicable to general linear systems.

The starting point for our theory is the ring-theoretic fractional representation theory introduced by the authors in a series of recent papers [9], [16] in which the set of compensators for a given plant are parameterized. Moreover, as a first cut at the simultaneous stabilization problem, one can reverse the role of the plant and compensator in this theory to parameterize the set of plants which are stabilized by a given compensator. In practice, however, one is not given a compensator *a priori* and, as such, we must characterize the set of plants obtained by the latter

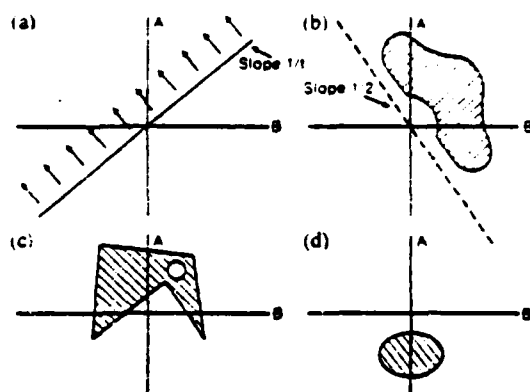


Fig. 1.

parameterization independently of the choice of compensator. For instance, in our first-order example, the set of plants in Fig. 1(a) are stabilized by a given compensator with slope $1/t$, while a given set of plants is simultaneously stabilizable if and only if it lies in the half-plane above some line through the origin. For the general problem, this is achieved by translating the fractional representation theory into an appropriate geometric setting in which the "shape" of the set of plants obtained from the latter parameterization may be characterized. In turn, the simultaneous stabilization problem may be resolved by requiring that the given set of plants lie in a region of the appropriate "shape."

Indeed, the appropriate geometric setting proves to be just the Grassmannian first introduced into the system theory literature by Hermann and Martin [11], [12]. Unlike their frequency domain formulation, however, we obtain the Grassmannian directly from the ring-theoretic fractional representation previously employed by the authors. Indeed, the Grassmannian is obtained simply by factoring out the nonuniqueness inherent in the fractional representation theory. As such, in addition to formulating the global theory necessary for our study of the simultaneous stabilization problem, the geometric approach yields new insight into the relationship between the fractional representation theory (which we identify with the elements of a general linear group) and the system itself (which we identify with the elements of a Grassmannian).

In the following section, the fractional representation theory is reviewed and the required Grassmannian is constructed. The resultant theory is then used to formulate a global description of the set of stabilizing compensators for a given plant in Section III. The resultant formulation also yields new insight into the problem of stabilizing a plant with a stable compensator [25] for which a necessary and sufficient condition is also derived in Section III. Finally, the simultaneous stabilization problem is investigated in Section IV wherein both global geometric and local algebraic criteria for the simultaneous stabilization of a prescribed family of plants are obtained.

Although the present paper is formulated in terms of the abstract (pseudo) coprime fractional representation theory of [5], [9], [10], [16]–[19], it should be pointed out that this

is simply one manifestation of a family of related approaches to the control system design problem developed during the past half-dozen years by Pernebo and Astrom [13], [14]; Antsaklis, Pearson, and Cheng [1], [6], [7]; Youla, Bongiorno, Jabr, and Lu [22]–[25]; and Zames [26] among others [2], [20], [21]. Indeed, the approaches of these authors are all closely related, any one of which could have been used as the basis for the present investigations. In particular, the formulation of Zames is applicable in a general ring-theoretic setting, and is essentially equivalent to that employed herein.

II. FRACTIONAL REPRESENTATION AND THE GRASSMANNIAN

The algebraic fractional representation theory is set in a nest of rings, groups, and multiplicative structures:

$$g \supset h \supset i \supset j.$$

Here, g is a ring with identity which represents the general class of systems with which we wish to work: rational matrices, continuous operators, a class of transcendental functions, etc.; and h is a subring of g containing the identity which models the systems which are stable in some sense: poles in a prescribed region, transcendental functions with restricted singularities, causal operators, etc. Finally, i denotes the multiplicative set composed of elements of h which admit an inverse in g , while j denotes the multiplicative subgroup of i made up of elements which are invertible in h . Detailed examples of this structure were given in [4] and [8] and will not be repeated here.

We say that a system $s \in g$ has a *right fractional representation* in $\{g, h, i, j\}$ if

$$s = n_r d_r^{-1} \quad (2.1)$$

where $n_r \in h$ and $d_r \in i$. Furthermore, we say that this representation is *right coprime* if there exist u_r and v_r in h such that

$$u_r n_r + v_r d_r = 1. \quad (2.2)$$

This equality is equivalent to the classical coprimeness concept for rational functions and matrices, while being defined in our general ring-theoretic setting. In particular, if g is the ring of rational functions and h is the ring of polynomials, (2.2) implies that n_r and d_r have no common zeros; and if g is the ring of rational functions and h is the ring of exponentially stable rational functions, (2.2) implies that n_r and d_r have no common right half-plane zeros.

Since g is, in general, noncommutative, we also define a *left fractional representation* for s via the equality

$$s = d_l^{-1} n_l \quad (2.3)$$

for $n_l \in h$ and $d_l \in i$. Furthermore, we say that this representation is *left coprime* if there exist u_l and v_l in h such that

$$n_l u_l - d_l v_l = 1. \quad (2.4)$$

Of course, in the classical case of a rational function or matrix, these fractional representations are assured to exist [10]. However, this is not the case in the general ring-theoretic setting. Therefore, for distributed, time-varying, and multidimensional systems, we assume that our plants admit such a representation as a prerequisite to the theory.¹ Interestingly, however, if such a representation exists, we may, without loss of generality, choose u_{sr} and v_{sl} such that the equality

$$v_{sr}u_{sl} = u_{sr}v_{sl} \quad (2.5)$$

also holds [8]. In this case, we say that the representation for s defined by the eight-tuple $\{n_{sr}, d_{sr}, u_{sr}, v_{sr}, n_{sl}, d_{sl}, u_{sl}, v_{sl}\}$ is *doubly coprime* and we express the defining equalities of (2.1)–(2.5) via the matrix equality $R_s^{-1} = S_s$ where

$$R_s = \begin{bmatrix} d_{sr} & -u_{sr} \\ n_{sr} & v_{sr} \end{bmatrix} \quad (2.6)$$

and

$$S_s = \begin{bmatrix} v_{sr} & u_{sr} \\ -n_{sl} & d_{sl} \end{bmatrix}. \quad (2.7)$$

It is interesting to compare the above formulation with that proposed by Zames [26]. Rather than working with an unstable system, s , Zames assumes that the given plant is first stabilized via classical techniques, and then develops his design theory around the resultant stable plant \tilde{s} . Now, since \tilde{s} is stable, it admits the *trivial right coprime representation* $\tilde{s} = \tilde{s}1^{-1}$ with the equality

$$[0][s] + [1][s] = u_{sr}n_{sr} + v_{sr}d_{sr} = 1 \quad (2.8)$$

implying right coprimeness, while a similar equality defines a left coprime representation for \tilde{s} . As such, the matrices $R_{\tilde{s}}$ and $S_{\tilde{s}}$ take on a very special form, permitting Zames to implement his design theory without explicitly dealing with u 's and v 's nor even introducing the coprimeness concept.

The key to our geometric formulation of the fractional representation theory lies with the observation that the 2×2 matrices R_s and S_s constitute a natural and concise representation for the given plant which can serve in lieu of the eight-tuple of n 's, d 's, u 's, and v 's. Indeed, if the input-output relation for our system is characterized by the equality

$$v_s = s\epsilon_s = [n_{sr}d_{sr}^{-1}]\epsilon_s, \quad (2.9)$$

then the admissible input-output pairs [9] (ϵ_s, v_s) for our plant are parameterized by the equality

$$\begin{bmatrix} \epsilon_s \\ v_s \end{bmatrix} = \begin{bmatrix} d_{sr} & -u_{sr} \\ n_{sr} & v_{sr} \end{bmatrix} \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = R_s \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} \quad (2.10)$$

¹Computationally, the evaluation of u_{sr} and v_{sr} (or u_{sl} and v_{sl}) reduces to the solution of a linear equation in the ring h . In particular, if h is a ring of stable rational functions or matrices thereof, one can multiply (2.2) and (2.4) by a common denominator, and thereby reduce their solution to that of a classical polynomial equation.

where $\sigma_s = [d_{sr}^{-1}]\epsilon_s$ is an appropriate "partial state" variable. As such, the 2×2 matrix R_s defines a natural model for the given system. Indeed, when such a model is employed, one can drop the invertibility requirement on d_{sr} (and d_{sl}), although the matrix R_s must still admit an inverse with entries in h .²

Since the 2×2 matrices R_s and S_s have entries in h and admit an inverse which also has entries in h , they form a group which we denote by $GL_h(2)$, i.e., the general linear group of 2×2 matrices with entries in h . If the elements of $GL_h(2)$ are, however, to serve as a viable system representation, we must be cognizant of the fact that several such matrices may represent the same plant. The appropriate equivalence classes may, however, be characterized with the aid of the subgroup $E \subset GL_h(2)$ composed of the upper triangular matrices

$$E = \begin{bmatrix} e_{11} & e_{12} \\ 0 & e_{22} \end{bmatrix} \quad (2.11)$$

in $GL_h(2)$. Note that since these triangular matrices are assumed to be in $GL_h(2)$, it follows that e_{11} and e_{22} are in j .

Property 1: Let R_s and \tilde{R}_s be in $GL_h(2)$. Then R_s and \tilde{R}_s represent the same system if and only if there exists $E \in E$ such that $R_s = \tilde{R}_s E$.

Proof: If $R_s = \tilde{R}_s E$ for some $E \in E$, then

$$\begin{bmatrix} \epsilon_s \\ v_s \end{bmatrix} = R_s \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = \tilde{R}_s E \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = \tilde{R}_s \begin{bmatrix} e_{11}\sigma_s \\ 0 \end{bmatrix} = \tilde{R}_s \begin{bmatrix} \tilde{\sigma}_s \\ 0 \end{bmatrix} \quad (2.12)$$

where $\tilde{\sigma}_s = e_{11}\sigma_s$. As such, the set of input-output pairs defined by \tilde{R}_s coincides with those defined by R_s , except for a change of parameterization.

If R_s and \tilde{R}_s define the same set of input-output pairs, then for any such pair (ϵ_s, v_s) , there exist σ_s and $\tilde{\sigma}_s$ such that

$$\begin{bmatrix} \epsilon_s \\ v_s \end{bmatrix} = R_s \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = \tilde{R}_s \begin{bmatrix} \tilde{\sigma}_s \\ 0 \end{bmatrix} \quad (2.13)$$

which, in turn, implies that

$$\begin{bmatrix} \tilde{\sigma}_s \\ 0 \end{bmatrix} = \tilde{R}_s^{-1} R_s \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = E \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix}. \quad (2.14)$$

Now, $E = \tilde{R}_s^{-1} R_s \in GL_h(2)$ since R_s and \tilde{R}_s are in $GL_h(2)$, while $R_s = \tilde{R}_s E$ by construction. As such, it suffices to show that $E \in E$. This, however, follows from the fact that (2.14) holds for all σ_s (and a corresponding $\tilde{\sigma}_s$). ■

Given that any two representations in $GL_h(2)$ for the same system differ by a left factor in E , a natural setting for our system theory is the quotient space $GL_h(2)/E$. Although $GL_h(2)$ is a group and E is a subgroup, E is not normal and, as such, $GL_h(2)/E$ is not a group. Fortunately, however, the resultant coset space (of equivalence classes) is a well-known and much studied geometric ob-

²In the case where d_{sr}^{-1} does not exist, the plant defines a relation rather than a function on the input-output space. The resultant relation is, however, parameterized by σ_s . Moreover, since R_s is invertible, the relation is normal in the same sense of [15].

ject, the Grassmannian [1] $G_k(1,2)$, which we will adopt as the basic setting for our system theory.³

Since E is not a normal subgroup of $GL_k(2)$, $G_k(1,2)$ is not a group and, as such, does not admit an "internal" algebraic structure. The group $GL_k(2)$ does, however, act as a set of transformations on $G_k(1,2)$. Indeed, if $T \in GL_k(2)$ and $[U] \in G_k(1,2)$ is the equivalence class of $U \in GL_k(2)$, we define

$$T[U] = [TU] \in G_k(1,2). \quad (2.15)$$

Now, if $[U] = [V]$, then Property 1 implies that there exists $E \in E$ such that $V = UE$; hence,

$$T[V] = T[UE] = [TUE] = [TU] \quad (2.16)$$

via Property 1. As such, the operation of $GL_k(2)$ on $G_k(1,2)$ is well defined.

As a prerequisite to the formulation of our stabilization theory, it is necessary to characterize the stable systems as interpreted in $GL_k(2)$ and $G_k(1,2)$. Recall that if s is stable ($s \in h$), then s admits the doubly coprime fractional representation

$$R_s = \begin{bmatrix} 1 & 0 \\ s & 1 \end{bmatrix}. \quad (2.17)$$

Denoting the set of such 2×2 matrices by $W \subset GL_k(2)$, it then follows from Property 1 that the set of all representations for the stable systems in $GL_k(2)$ take the form $S = WE$ and, as such, they are represented by $\{S\} = \{WE\} = [W] \subset G_k(1,2)$. Although W and E are both subgroups of $GL_k(2)$, they are not commutative and, as such, $S = WE$ is not a group. S is, however, characterized by the following property.

Property 2: Let

$$T = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}$$

be in $GL_k(2)$. Then $T \in S$ if and only if $t_{11} \in j$.

Proof: Since $S = WE$ if $T \in S$, then $T = WE$ for some $W \in W$ and $E \in E$; hence,

³If h is the field of scalars, $G_k(1,2)$ is the classical Grassmannian (of lines in 2-space), whereas if h is taken to be $n \times n$ matrices, $G_k(1,2)$ reduces to the classical Grassmannian of n planes in $2n$ space [1]. Although these classical Grassmannians have an analytic structure (compact manifold) which may not be shared by our abstract Grassmannian, the algebraic properties of $G_k(1,2)$ are all that is required for the present theory, and hence no difficulty is encountered by working with elements taken from an abstract ring. Indeed, for our purposes, we only use the fact that $G_k(1,2)$ is the coset space $GL_k(2)/E$ and identify it as the Grassmannian only to make the connection with the classical literature.

It is interesting to note that the Grassmannian has been used as a natural setting for multivariate systems by mathematical system theorists for a number of years [28], [29], [32]. Here, a system represented by an $n \times n$ frequency response matrix is identified with a curve taking values in the classical Grassmannian of n planes in $2n$ space. As such, that theory identifies a system with a Grassmannian-valued function, while our formulation identifies the system with a Grassmannian built from a ring of functions. Of course, the two approaches are completely equivalent in the multivariate case, while the present formulation is also well defined for general linear systems (time-varying, distributed, etc.)

$$\begin{aligned} T &= \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ w & 1 \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} \\ 0 & e_{22} \end{bmatrix} \\ &= \begin{bmatrix} e_{11} & e_{12} \\ we_{11} & (we_{12} + e_{22}) \end{bmatrix} \end{aligned} \quad (2.18)$$

showing that $t_{11} = e_{11} \in j$ [since E is invertible in $GL_k(2)$]. Conversely, if $t_{11} \in j$, we may factor T via

$$T = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ t_{21}t_{11}^{-1} & 1 \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} - t_{21}t_{11}^{-1}t_{12} \end{bmatrix} = WE. \quad (2.19)$$

Since $t_{11} \in j$, $w = t_{21}t_{11}^{-1}$ and $t_{22} - t_{21}t_{11}^{-1}t_{12}$ are in h . Thus, the unit triangular nature of W implies that it is in $GL_k(2)$. This, in turn, however, implies that

$$E = W^{-1}T \in GL_k(2) \quad (2.20)$$

since $GL_k(2)$ is a group. Since E is upper triangular and $E \in GL_k(2)$, $E \in E$, as was to be shown. ■

Since W and E are both groups, the equality $S = WE$ implies that $S^{-1} = EW$. Now, a little algebra similar to that used in the proof of Property 2 will reveal that $T \in S^{-1}$ if and only if $t_{22} \in j$. Indeed, S and S^{-1} are precisely the two classes of matrices for which the 2×2 matrix inversion formula is applicable [16].

Before concluding this section, it is instructive to comment on the relationship between $GL_k(2)$ and $G_k(1,2)$. In essence, $GL_k(2)$ is a set of representations for our systems, while $G_k(1,2)$ represents the set of systems. That is to say, $GL_k(2)$ is composed of the computationally tractable objects with which we actually describe a system, although many such objects may represent the same system. On the contrary, each element of $G_k(1,2)$ is uniquely identified with a system, and hence we may think of $G_k(1,2)$ as "being the set of systems" (or, at least, being in one-to-one correspondence with the set of systems). On the other hand, the elements of $G_k(1,2)$ are not computationally tractable, except through the intermediary of $GL_k(2)$. In practice, therefore, a plant is characterized by one of its representations in $GL_k(2)$, while the goal of the system design problem is to specify an appropriate compensator in $G_k(1,2)$. That is, we are designing a compensator rather than a representation of a compensator, and hence even though we work in $GL_k(2)$ as a matter of computational necessity, the result of the design process is an element of $G_k(1,2)$.

III. STABILIZATION

The basic feedback system we consider is shown in Fig. 2. The system is characterized by the connection equations

$$e_p = \mu_z - v_c \quad (3.1)$$

and

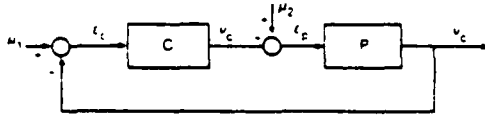


Fig. 2.

$$\epsilon_c = \mu_1 - v_p, \quad (3.2)$$

while the plant and compensator are characterized by

$$\begin{bmatrix} \epsilon_p \\ v_p \end{bmatrix} = R_p \begin{bmatrix} \sigma_p \\ 0 \end{bmatrix} \quad (3.3)$$

and

$$\begin{bmatrix} \epsilon_c \\ v_c \end{bmatrix} = R_c \begin{bmatrix} \sigma_c \\ 0 \end{bmatrix}, \quad (3.4)$$

respectively, where R_p and R_c are in $GL_n(2)$. Letting Q be the 2×2 matrix in $GL_n(2)$ defined by

$$Q = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad (3.5)$$

the connection equations (3.1) and (3.2) can be expressed as

$$\begin{bmatrix} \epsilon_p \\ v_p \end{bmatrix} = Q \begin{bmatrix} \epsilon_c \\ v_c \end{bmatrix} + \begin{bmatrix} \mu_2 \\ \mu_1 \end{bmatrix}. \quad (3.6)$$

A little algebra will also reveal that $Q^{-1} = Q' = -Q$; hence, this connection matrix is readily manipulated. Finally, the substitution of (3.3) and (3.4) into (3.6) yields the equality

$$\begin{aligned} \begin{bmatrix} \mu_2 \\ \mu_1 \end{bmatrix} &= R_p \begin{bmatrix} \sigma_p \\ 0 \end{bmatrix} - Q R_c \begin{bmatrix} \sigma_c \\ 0 \end{bmatrix} = R_p P_1 \begin{bmatrix} \sigma_p \\ \sigma_c \end{bmatrix} - Q R_c P_2 \begin{bmatrix} \sigma_p \\ \sigma_c \end{bmatrix} \\ &= [R_p P_1 + Q' R_c Q P_2] \begin{bmatrix} \sigma_p \\ \sigma_c \end{bmatrix}. \end{aligned} \quad (3.7)$$

Here $P_1 = \text{diag}[1, 0]$ and $P_2 = \text{diag}[0, 1]$.

On the basis of the above formulation, we say that the system is *stable* if and only if

$$[R_p P_1 + Q' R_c Q P_2] \in GL_n(2). \quad (3.8)$$

Since $[R_p P_1 + Q' R_c Q P_2]$ has entries in \mathcal{H} , (3.8) implies that its inverse exists and also has entries in \mathcal{H} . As such, a feedback system will be stable if and only if the relationship between its input vector $\text{col}(\mu_2, \mu_1)$ and its partial state vector $\text{col}(\sigma_p, \sigma_c)$ is stable. With the aid of (3.3) and (3.4), this, in turn, implies that the relationship between the system input vector and all of its internal variables is stable, and the converse is also true [9].

With these preliminaries, we now proceed with our first theorem in which a geometric characterization of the set of stabilizing compensators for a given plant is obtained. To this end, recall that S is the subset of $GL_n(2)$ corresponding to the stable systems and $[S]$ is its image in $G_n(1, 2)$, while any $T \in GL_n(2)$ defines a transformation on $[S] \subset G_n(1, 2)$ via

$$T[S] = \{[TS] : S \in S\} \subset G_n(1, 2). \quad (3.9)$$

Finally, let $R_c(R_p)$ denote the set of stabilizing compensators in $G_n(1, 2)$ for a given plant represented by $R_p \in GL_n(2)$.

Theorem: $R_c(R_p) = QR_p Q'[S]$.

Proof: To prove the theorem, we will show that the set of all representations for the stabilizing compensators in $GL_n(2)$ takes the form $QR_p Q'S$ for some $S \in S$, from which the theorem follows upon mapping these presentations into corresponding systems which are identified as elements of $G_n(1, 2)$. If $R_c = QR_p Q'S$ for some $S \in S$, then

$$\begin{aligned} [R_p P_1 + Q' R_c Q P_2] &= [R_p P_1 + Q'(QR_p Q'S)Q P_2] \\ &= R_p [P_1 + Q'SQ P_2]. \end{aligned} \quad (3.10)$$

Now, if

$$S = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \quad (3.11)$$

with $s_{11} \in j$, since $S \in S$, a little algebra will reveal that

$$Q'SQ = \begin{bmatrix} s_{22} & -s_{21} \\ -s_{12} & s_{11} \end{bmatrix} \in S^{-1} \quad (3.12)$$

since $(Q'SQ)_{22} = s_{11} \in j$. As such,

$$R_p [P_1 + Q'SQ P_2] = R_p \begin{bmatrix} 1 & -s_{21} \\ 0 & s_{22} \end{bmatrix} \in GL_n(2). \quad (3.13)$$

showing that the feedback system with compensator $QR_p Q'S$ is stable.

Conversely, if R_c is a stabilizing compensator for the feedback system,

$$[R_p P_1 + Q' R_c Q P_2] \in GL_n(2). \quad (3.14)$$

Moreover, since Q and R_p are in $GL_n(2)$, we may, without loss of generality, assume that R_c is of the form

$$R_c = QR_p Q'X, \quad (3.15)$$

in which case it suffices to show that

$$X = QR_p^{-1} Q' R_c \in S. \quad (3.16)$$

To this end, we observe that

$$\begin{aligned} [R_p P_1 + Q' R_c Q P_2] &= [R_p P_1 + Q'(QR_p Q'X)Q P_2] \\ &= R_p [P_1 + Q'XQ P_2] \\ &= R_p \begin{bmatrix} 1 & -x_{21} \\ 0 & x_{11} \end{bmatrix} \end{aligned} \quad (3.17)$$

since

$$X'XQ = \begin{bmatrix} x_{22} & -x_{21} \\ -x_{12} & x_{11} \end{bmatrix}. \quad (3.18)$$

Now, since $[R_p P_1 + Q' R_c Q P_2] \in GL_k(2)$,

$$\begin{bmatrix} 1 & -x_{21} \\ 0 & x_{11} \end{bmatrix} = R_p^{-1} [R_p P_1 + Q' R_c Q P_2] \in GL_k(2), \quad (3.19)$$

which implies that $x_{11} \in j$ as required to verify that $X \in S$. As such, an arbitrary stabilizing compensator for our system is of the form $R_c = Q R_p Q' X$.

The set of all representations of all stabilizing compensators in $GL_k(2)$ thus takes the form $Q R_p Q' S \subset GL_k(2)$. Now, upon mapping this set into the Grassmannian, we obtain

$$[Q R_p Q' S] = Q R_p Q' [S], \quad (3.20)$$

completing the proof. ■

Unlike the previous results given directly in terms of the fractional representation theory [9], [24] which are local in nature, the present theorem yields a global description of the set of stabilizing compensators for a given plant. Indeed, the required set is just a copy of the stable system in the Grassmannian transformed by the action of $Q R_p Q'$.

It is interesting to compare the parameterization of the stabilizing compensators of the theorem with that obtained directly from the fractional representation theory. Indeed, a little algebra with the results of [4] will yield the equality

$$R_c = Q R_p Q' W; \quad W \in W \quad (3.21)$$

for the family of stabilizing compensators. Recalling, however, that $S = WE$, this parameterization differs from that of the theorem only in that the equivalence transformation E has been deleted. As such, the set of stabilizing compensators of (3.21) includes exactly one representation for each stabilizing compensator rather than parameterizing all representations for the stabilizing compensators as is the case with the present theorem. Of course,

$$[Q R_p Q' W] = Q R_p Q' [W] = Q R_p Q' [S], \quad (3.22)$$

showing that the same set of compensators in the Grassmannian is defined by the two theories.

It follows immediately from the theorem that a stabilizing compensator always exists given that the plant is modeled by an $R_p \in GL_k(2)$. In practice, however, one often requires that the compensator be of a special form: stable, memoryless, diagonal, etc. As such, if $[C]$ represents the desired class of compensators in $G_k(1,2)$, it is necessary to design a compensator in $[C] \cap Q R_p Q' [S]$, in which case it is not clear that such a compensator even exists. In the case where $[C]$ represents the stable systems, i.e., we desire a stable stabilizing compensator [25], a simple method for the existence of the required compensator can be obtained as follows.

Lemma 1: A plant represented by $R_p \in GL_k(2)$ can be stabilized by a stable compensator if and only if

$$R_p \in S^{-1} S.$$

Proof: If $R_p \in S^{-1} S$, there exist S_1 and S_2 in S such that

$$R_p = S_2^{-1} S_1. \quad (3.23)$$

Now let

$$R_c = Q S_2^{-1} Q' \quad (3.24)$$

which is stable since the congruence transformation defined by Q maps S^{-1} to S and conversely. Moreover, since $Q' = Q^{-1}$,

$$\begin{aligned} R_c &= Q S_2^{-1} Q' = Q S_2^{-1} Q' Q S_1 Q' Q S_1^{-1} Q' \\ &= (Q S_2^{-1} S_1 Q') (Q S_1^{-1} Q') = Q R_p Q' S \end{aligned} \quad (3.25)$$

where $S = Q S_1^{-1} Q' \in S$. As such, we have constructed a stable stabilizing compensator for R_p as required.

Conversely, if $R_c \in S$ is a stable stabilizing compensator for R_p , then by the theorem, $R_c = Q R_p Q' S$ for some $S \in S$. As such,

$$\begin{aligned} R_p &= Q' R_c S^{-1} Q = Q' R_c Q Q' S^{-1} Q \\ &= (Q' R_c Q) (Q' S^{-1} Q) = S_2^{-1} S_1 \end{aligned} \quad (3.26)$$

since the congruence transformation defined by Q maps $R_c \in S$ to $S_2^{-1} \in S^{-1}$ and $S^{-1} \in S^{-1}$ to $S_1 \in S$. ■

Of course, when the hypotheses of the corollary hold, the set of stable stabilizing compensators for $R_p = S_2^{-1} S_1$ is nonvoid and given by $[S] \cap Q R_p Q' [S]$. No explicit parameterization for this set is, however, known nor is it obvious that one even exists. This intuition will be explored further in the following section on the simultaneous stabilization problem.

IV. SIMULTANEOUS STABILIZATION

The key to our solution of the simultaneous stabilization problem lies with a reversal of the analysis used in the derivation of the stabilization theorem. That is, one assumes that $R_c \in GL_k(2)$ is given and parameterizes the set of plants which are stabilized by R_c . Denoting that set of plants by $R_p(R_c)$, we obtain the following theorem. Since the proof for the theorem is virtually identical to that already given for the stabilization theorem of the previous section, it will not be repeated here.

Theorem: $R_p(R_c) = Q R_c Q' [S]$.

Since every $T \in GL_k(2)$ is of the form $T = Q R_c Q'$ (with $R_c = Q' T Q$), the lemma implies that a set of plants $P \subset G_k(1,2)$ can be simultaneously stabilized if and only if they lie in a copy of the stable systems $[S]$ transformed by the action of some $T \in GL_k(2)$. Indeed, this is the desired generalization of the first-order example given in the introduction to the case of a general linear system. In the general case, the Grassmannian plays the role of the A - B plane, the stables, $[S]$, play the role of the half-plane, and the general linear group, $GL_k(2)$, serves as the group of "rotations." These observations are summarized in the following corollary.

Corollary 2: A set of plants $P \subset G_k(1,2)$ can be simultaneously stabilized if and only if they lie in $T[S]$ for some $T \in GL_k(1,2)$.

Although the corollary represents a complete geometric solution to the simultaneous stabilization problem, it is not amenable to convenient implementation. We therefore give an alternative algebraic solution to the problem in $GL_n(2)$.

Corollary 3: Let $R_p \in GL_n(2)$, $p \in P$ be representations for a set of plants $P \subset G_n(1,2)$. Then P can be simultaneously stabilized if and only if there exists a family of matrices $S_p \in S$, $p \in P$ such that

$$S_p^{-1}S_q = R_p^{-1}R_q; \quad p, q \in P.$$

Proof: If R_c simultaneously stabilizes R_p , $p \in P$, then it follows from Corollary 2 that there exist $S_p \in S$, $p \in P$ such that

$$R_p = QR_cQ'S_p; \quad p \in P. \quad (4.1)$$

Hence,

$$R_p^{-1}R_q = (QR_cQ'S_p)^{-1}(QR_cQ'S_q) = S_p^{-1}S_q. \quad (4.2)$$

Conversely if there exists $S_p \in S$, $p \in P$ such that $R_p^{-1}R_q = S_p^{-1}S_q$; $p, q \in P$, then we let p' be an arbitrary plant in P and define R_c via

$$R_c = Q'R_pS_p^{-1}Q. \quad (4.3)$$

R_c is independent of the choice of p' . Indeed, if p'' is an alternative choice, then since $R_{p'}^{-1}R_{p''} = S_{p'}^{-1}S_{p''}$,

$$R_{p'}S_p^{-1} = R_{p'}(R_{p'}^{-1}R_{p''}S_{p''}^{-1}) = R_{p''}S_{p''}^{-1}. \quad (4.4)$$

Moreover, for any $p \in P$, $R_{p'}^{-1}R_p = S_{p'}^{-1}S_p$; hence,

$$\begin{aligned} R_p &= R_{p'}S_p^{-1}S_p = QQ'R_{p'}S_p^{-1}QQ'S_p \\ &= Q(Q'R_{p'}S_p^{-1}Q)Q'S_p = QR_cQ'S_p \end{aligned} \quad (4.5)$$

from which it follows, via the corollary, that the set of plants $P \subset G_n(1,2)$ is simultaneously stabilized by R_c . ■

It is interesting to note that in the special case in which P contains exactly two plants, say p and q , then they are simultaneously stabilizable if and only if

$$R_p^{-1}R_q \in S^{-1}S \quad (4.6)$$

which is identical to the criterion of Corollary 1 to stabilize a plant with a stable compensator. Indeed, this should not be surprising since the problem of stabilizing a plant with a stable compensator is completely equivalent to the problem of simultaneously stabilizing the given plant and the zero plant [represented by $R_z = 1 \in GL_n(2)$]. This follows immediately from the definition of stability used in [4] or, alternatively, one may let $R_z = 1$ and

$$R_c = \begin{bmatrix} d_{cr} & -u_{cl} \\ n_{cr} & v_{cl} \end{bmatrix} \quad (4.7)$$

in which case

$$[R_zP_1 + Q'R_cQP_2] = \begin{bmatrix} 1 & -n_{cr} \\ 0 & d_{cr} \end{bmatrix}. \quad (4.8)$$

As such, R_c stabilizes the zero plant if and only if $d_{cr} \in j$ or, equivalently, $R_c \in S$. The problem of stabilizing p with a stable compensator is thus equivalent to simultaneously stabilizing p and the zero plant. Since the zero plant is represented by $1 \in GL_n(2)$, it follows from (4.6) that p can be stabilized by a stable compensator if and only if

$$R_p = (1)^{-1}R_p = R_c^{-1}R_p \in S^{-1}S \quad (4.9)$$

which replicates the result of Corollary 3.

Of course, a similar argument can be used to obtain an algebraic criterion for the simultaneous stabilization of a set of plants P by a stable compensator. To this end, we simply augment the set of plants P by the zero plant and then apply the theorem to $P' = \{0\} \cup P$. Alternatively, a geometric criterion for the simultaneous stabilization of P may be obtained by requiring that $P \subset T[S]$ for some $T \in S^{-1}$. Since the corresponding R_c is given by

$$R_c = Q'TQ, \quad (4.10)$$

the resultant compensator will be stable.

Finally, the possibility of parameterizing the set of compensators (or stable compensators) which simultaneously stabilize P should be considered. In effect, this amounts to parameterizing the set of 2×2 matrices $T \in GL_n(2)$ or $T \in S^{-1}$ such that $P \subset T[S]$ which, in turn, requires some kind of parameterization for P . In particular, if $P = T[S]$, then $R_c = Q'TQ$ is the unique stabilizing compensator, while the stabilizing compensators for a single plant p may be parameterized by $[S]$ as per the stabilization theorem of Section III.

V. EXAMPLES

Since the action of $GL_n(2)$ on $G_n(1,2)$ is a geometric invariant, one can, at least intuitively, say that the "shape" of $T[S]$ is identical to that of $[S]$, and every set whose "shape" is the same as $[S]$ may be obtained from $[S]$ by such a transformation. As such, Corollary 4 implies that a prescribed set of plants $P \subset G_n(1,2)$ admits a simultaneous stabilization if and only if P is contained in a subset of the Grassmannian whose "shape" is the same as $[S]$. For instance, in the example of Fig. 1, P must be contained in an appropriate half-plane.

Although well-defined coordinate systems do exist in the Grassmannian, at the time of this writing we have yet to formulate a computational algorithm for implementing the above-described simultaneous stabilization problem in any degree of generality. Interestingly, however, in the case where P is composed of exactly two plants (a two-speed motor or the dynamics of a swing-wing aircraft), a simple frequency domain criterion for the simultaneous stabilization problem was given by Chua [8], [16] and the authors in the single-variate case, and has since been extended to the multivariate case by Vidyasagar and Viswanadham [19].

To illustrate the above-described general geometric criterion in a computationally trackable setting, consider the

case of a set of third-order single-variate plants

$$p(s) = \frac{q_2 s^2 + q_1 s + q_0}{s^3 + p_2 s^2 + p_1 s + p_0} \quad (5.1)$$

which are to be stabilized by a proportional compensator $c(s) = t$. To represent the class of plants geometrically, we identify the six-dimensional space of such plants with a family of three-dimensional Euclidian spaces (with coordinates p_0, p_1, p_2) parameterized by the numerator coefficients q_0, q_1 , and q_2 . As such, we have a three-dimensional family of three-dimensional spaces $R^3_{(q_0, q_1, q_2)}$. Of course, if there is no feedback ($t = 0$), such a system will be stable if and only if

$$p_0, p_2 > 0 \quad (5.2)$$

and

$$p_1 > p_0/p_2 \quad (5.3)$$

independently of the numerator parameters. The resultant stability region for the $t = 0$ case is thus as illustrated in Fig. 3(a).

In the case of a nonzero compensator, a similar argument with the Hurwitz criterion will yield the set of inequalities

$$p_0 + tq_0 > 0 \quad (5.4)$$

$$p_2 + tq_2 > 0 \quad (5.5)$$

and

$$p_1 > (p_0 + tq_0)/(p_2 + tq_2). \quad (5.6)$$

As such, for a fixed value of the numerator coefficients q_0, q_1 , and q_2 , the new stability region is identical in "shape" to the $t = 0$ case except for a shift of the origin to the point (tq_0, tq_1, tq_2) in $R^3_{(q_0, q_1, q_2)}$, as illustrated in Fig. 3(b). Of course, as t changes, the origin of the stability region moves along the line determined by the point (q_0, q_1, q_2) . Unlike the $t = 0$ case, however, where the stability region is independent of the numerator parameters, the line along which the origin of the stability region moves is determined by the numerators parameters. Thus, in this example, our three-dimensional family of three-dimensional Euclidian spaces $R^3_{(q_0, q_1, q_2)}$ plays the role of the *Grassmannian*, with the region defined by (5.2) and (5.3) in each such space characterizing the *stable systems*. Furthermore, the *real group* (corresponding to the proportional compensators) acts on this space by translating the stability region in the space $R^3_{(q_0, q_1, q_2)}$ along a line determined by the point (q_0, q_1, q_2) .

Although the above example was derived from basic principles, we believe that it illustrates the essential geometric nature of the simultaneous stabilization problem as formulated in our abstract theory. Indeed, a prescribed set of plants can be simultaneously stabilized if and only if they are contained in a translate of the stables.

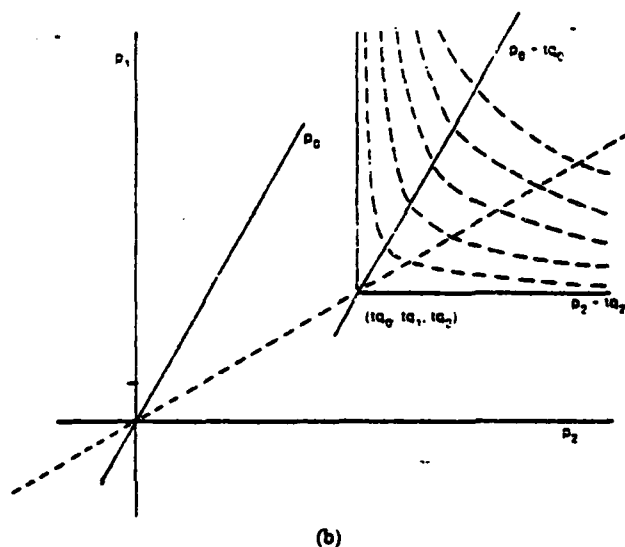
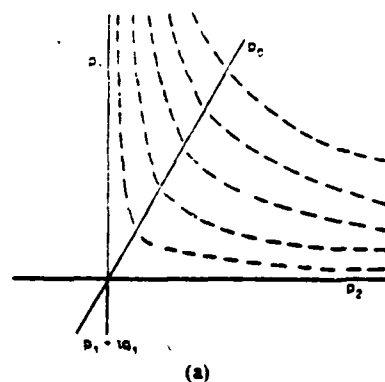


Fig. 3.

VI. CONCLUSIONS

Our purpose in the preceding has been threefold. First, we have attempted to exhibit the essential relationship between the fractional representation theory and the algebraic-geometric approach to system theory. Second, we have presented a global solution to the feedback system stabilization problem. Third, a solution to the simultaneous stabilization problem has been presented. It should, however, be pointed out that the solution presented for the simultaneous stabilization problem is mathematical in nature and not intended for computational implementation. At the present time, no computationally feasible solution to the simultaneous stabilization problem is known, except in the case where P contains exactly two plants wherein a simple frequency domain test is possible [8], [16] and in the simple low dimensional illustrated herein.

REFERENCES

- [1] P. J. Antsaklis and J. B. Pearson, "Stabilization and regulation in linear multivariable systems," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 928-930, 1978.
- [2] G. Bengtsson, "Output regulation and internal modes—A

- frequency domain approach." *Automatica*, vol. 13, pp. 333-345, 1977.
- [3] F. Brickell and R. S. Clark, *Differentiable Manifolds*. London, England: Van Nostrand Reinhold, 1970.
 - [4] R. W. Brockett and C. I. Byrnes, "Multivariable Nyquist criteria, root locus, and pole placement: A geometric viewpoint." *IEEE Trans. Automat. Contr.*, to be published.
 - [5] F. M. Callier and C. A. Desoer, "Stabilization, tracking and disturbance rejection in linear multivariable distributed systems." in *Proc. 17th IEEE Conf. Decision Contr.*, San Diego, CA, Jan. 1979, pp. 513-514; also, Univ. California, Berkeley, Tech. Memo. UCB/ERL M78/83, Dec. 1978.
 - [6] L. Cheng and J. B. Pearson, "Frequency domain synthesis of multivariable linear regulators." *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 3-15, 1978.
 - [7] —, "Synthesis of linear multivariable regulators." *IEEE Trans. Automat. Contr.*, to be published.
 - [8] O. Chua, M.S. thesis, Texas Tech Univ., Lubbock, 1980.
 - [9] C. A. Desoer, R.-W. Liu, J. Murray, and R. Saeks, "Feedback system design: The fractional representation approach to analysis and synthesis." *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 399-412, 1980.
 - [10] B. Francis and M. Vidyasagar, "Algebraic and topological aspects of the servo problem for lumped linear systems," unpublished notes, Yale Univ., New Haven, CT, 1980.
 - [11] R. Hermann and C. Martin, "Applications of algebraic geometry to linear system theory," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 19-25, 1977.
 - [12] —, "Applications of algebraic geometry to system theory: The MacMillan degree and Kronecker indices as topological and holomorphic invariants." *SIAM J. Contr.*, vol. 16, pp. 743-755, 1978.
 - [13] L. Pernbo, Ph.D. dissertation, Lund Inst. Technol., Lund, Sweden, 1978.
 - [14] —, "An algebraic theory for design of controllers for linear multivariable systems: Parts I and II," *IEEE Trans. Automat. Contr.*, to be published.
 - [15] R. Saeks, *Generalized Networks*. New York: Holt, Rinehart and Winston, 1972.
 - [16] R. Saeks, J. J. Murray, O. Chua, and C. Karmokolias, "Feedback system design: The single variate case," unpublished notes, Texas Tech Univ., Lubbock, 1980.
 - [17] R. Saeks and J. Murray, "Feedback systems design: The tracking and disturbance rejection problems," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 203-217, 1981.
 - [18] M. Vidyasagar, H. Schneider, and B. Francis, "Algebraic and topological aspects of feedback system stabilization," Dep. Elec. Eng., Univ. Waterloo, Waterloo, Ont., Canada, Tech. Rep. 80-09, 1980.
 - [19] M. Vidyasagar and M. Viswanadham, "Algebraic design techniques for reliable stabilization," Dep. Elec. Eng., Univ. Waterloo, Waterloo, Ont., Canada, Tech. Rep. 81-02, 1980.
 - [20] W. A. Wolovich, "Multivariable system synthesis with step disturbance rejection." *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 127-130, 1974.
 - [21] W. A. Wolovich and P. Ferreira, "Output regulation and tracking in linear multivariable systems." *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 460-465, 1979.
 - [22] D. C. Youla, "Interpolary multichannel spectral estimation," unpublished notes, Polytechnic Inst. New York, Brooklyn, 1979.
 - [23] D. C. Youla, J. J. Bongiorno, and H. A. Jabr, "Modern Wiener-Hopf design of optimal controllers, Part I." *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 3-15, 1976.
 - [24] —, "Modern Wiener-Hopf design of optimal controllers, Part II," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 319-338, 1976.
 - [25] D. C. Youla, J. J. Bongiorno, and C. N. Lu, "Single-loop feedback stabilization of linear multivariable dynamic plants," *Automatica*, vol. 10, pp. 159-173, 1974.
 - [26] G. Zames, "Feedback and optimal sensitivity: Model reference transformation, weighted seminorms, and approximate inverses," *IEEE Trans. Automat. Contr.*, to be published.



Richard Saeks (S'59-M'65-SM'74-F'77) was born in Chicago, IL, in 1941. He received the B.S. degree in 1964, the M.S. degree in 1965, and the Ph.D. degree in 1967 from Northwestern University, Evanston, IL, Colorado State University, Fort Collins, and Cornell University, Ithaca, NY, respectively, all in electrical engineering.

He is presently Paul Whitfield Horn Professor of Electrical Engineering, Mathematics, and Computer Science, Texas Tech University, Lubbock, where he is involved in teaching and research in the areas of fault analysis, large-scale systems, and mathematical system theory.

Dr. Saeks is a member of AMS, SIAM, ASEE, and Sigma Xi.



John Murray (M'78) was born in Galway, Ireland, on August 8, 1947. He received the B.Sc. and M.Sc. degrees from University College, Cork, Ireland, in 1969 and 1970, respectively, and the Ph.D. degree from the University of Notre Dame, Notre Dame, IN, in 1974, all in mathematics.

He is currently with the Department of Electrical Engineering, Texas Tech University, Lubbock. His principal research interests are in the areas of several complex variables, multidimensional system theory, and time-varying systems.

NONLINEAR CONTROL

8.4

THEORY OF DESIGN USING NONLINEAR TRANSFORMATIONS

Renjeng Su,* L. R. Hunt,† and George Meyer‡

NASA Ames Research Center
Moffett Field, California 94035

ABSTRACT

This paper is presenting an overview of the theory of transformations from nonlinear systems to linear systems. Topics covered include (1) necessary and sufficient conditions for transformations to exist, (2) a method of constructing transformations, (3) robustness in design (based on transformation theory) and Lyapunov functions, (4) estimation theory, and (5) the relationship between transformation theory and "nonlinear zeroes." Application of these results to automatic flight control is presented in another paper at this session.

I. INTRODUCTION

Suppose we have a system that is nonlinear. If there is a set of coordinates (in state and control space) for which this system appears as a linear system, then the nonlinearity is only the result of an unfortunate choice of coordinates. This leads us to the problem of classifying all nonlinear systems that can be transformed to controllable linear systems.

The systems of interest to us are

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^m u_i(t) g_i(x(t)), \quad (1)$$

where f, g_1, g_2, \dots, g_m are \mathcal{C}^∞ vector fields on \mathbb{R}^n [i.e., each assigns an n vector at points of \mathbb{R}^n] and $f(0) = 0$. We present necessary and sufficient conditions for these systems to be transformable to controllable linear systems in Brunovsky¹ canonical form. These conditions depend on Lie brackets from the field of differential geometry.

Given a transformable system, we mention a method for constructing a transformation to a linear system. This method depends on solving

*Research Associate of National Research Council at Ames Research Center.

†Research supported by Ames Research Center under the IPA program and the Joint Services Electronics Program at Texas Tech U. under Contract N00014-76-1136.

‡Research Engineer.

systems of ordinary differential equations. Several techniques for obtaining exact and approximate transformations are noted.

Once a nonlinear system is mapped to a linear system, the known theory for design of linear control systems can be used to design for the nonlinear system. We show that this technique is robust. In terms of stability, Lyapunov functions for the linear system are composed with the transformation to provide Lyapunov functions for the nonlinear system.

Brief comments are given showing the relationships between transformation theory and estimation and between transformation theory and "nonlinear zeroes."

For other work on transformations we refer to the research of Krener,² Brockett,³ Jakubczyk and Respondek,⁴ Hermann,⁵ and the results in Refs. 6-12.

II. TRANSFORMATIONS

If we have a controllable linear system with m inputs

$$\dot{y} = Ay + Bu \quad (2)$$

then Brunovsky¹ proved that it could be put in the canonical form

$$\dot{y} = A_0 y + E_0 v, \quad (3)$$

which is based on the Kronecker indices

$\kappa_1, \kappa_2, \dots, \kappa_m$ with $\kappa_1 + \kappa_2 + \dots + \kappa_m = n$ and $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_m$. We let $\sigma_1 = \kappa_1, \sigma_2 = \kappa_1 + \kappa_2, \dots, \sigma_m = \kappa_1 + \kappa_2 + \dots + \kappa_m = n$. For a complete description of Eq. (3) see Ref. 12.

We find necessary and sufficient conditions to map our nonlinear system (1) to the canonical linear system (3). We consider transformations $T = (T_1, T_2, \dots, T_n, T_{n+1}, \dots, T_{n+m})$ from $\mathbb{R}^{n+m}((x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m)$ space) to $\mathbb{R}^{n+m}((y_1, y_2, \dots, y_n, v_1, v_2, \dots, v_m)$ space) which apply for all (x_1, x_2, \dots, x_n) in some open neighborhood of $(0, 0, \dots, 0)$ in \mathbb{R}^n and which map the origin in x -space to the origin in y -space. Of course we want such transformations to be diffeomorphisms (\mathcal{C}^∞ , \mathcal{C}^∞ inverse, nonsingular, and one-one) and take system (1) to the system (3). The last m coordinate functions $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ are the

only ones that depend on the controls u_1, u_2, \dots, u_m .

Given \mathcal{W} vector fields f and g on R^n we define the Lie bracket of f and g :

$$[f, g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g,$$

where $\frac{\partial g}{\partial x}$ and $\frac{\partial f}{\partial x}$ are Jacobian matrices. Similarly we take

$$(\text{ad}^0 f, g) = g$$

$$(\text{ad}^1 f, g) = [f, g]$$

$$(\text{ad}^2 f, g) = [f, [f, g]]$$

...

$$(\text{ad}^k f, g) = [f, (\text{ad}^{k-1} f, g)]$$

A set of \mathcal{W} vector fields $\{f_1, f_2, \dots, f_r\}$ on R^n is involutive if there exist \mathcal{W} functions $\gamma_{ijk}(x)$ so that

$$[f_i, f_j](x) = \sum_{k=1}^r \gamma_{ijk}(x) f_k(x), \quad 1 \leq i, j \leq r, i \neq j.$$

We define the sets

$$C = \{g_1, [f, g_1], \dots, (\text{ad}^{\kappa_1-1} f, g_1), g_2, [f, g_2], \dots, (\text{ad}^{\kappa_2-1} f, g_2), \dots, g_m, [f, g_m], \dots, (\text{ad}^{\kappa_m-1} f, g_m)\}$$

$$C_j = \{g_1, [f, g_1], \dots, (\text{ad}^{\kappa_j-2} f, g_1), g_2, [f, g_2], \dots, (\text{ad}^{\kappa_j-2} f, g_2), \dots, g_m, [f, g_m], \dots, (\text{ad}^{\kappa_j-2} f, g_m)\} \text{ for } j = 1, 2, \dots, m$$

The following result from Ref. 11 gives conditions under which a transformation of the type we consider exists.

Theorem 1 The system (1) is transformable to the system (3), where the state variables x_1, x_2, \dots, x_n are contained in some open neighborhood V of the origin in R^n , if and only if

1. The set C spans an n -dimensional space at each point of V
2. The sets C_j are involutive for $j = 1, 2, \dots, m$
3. The span of each C_j is equal to the span of $C_j \cap C$

An obvious problem is how to construct a transformation if the hypotheses of this theorem are satisfied.

III. CONSTRUCTION OF TRANSFORMATIONS

We denote by $\langle \dots \rangle$ the duality between one-forms and vector fields; that is, if h is a \mathcal{W} -function and f a \mathcal{W} -vector field

$$\langle dh, f \rangle = \frac{\partial h}{\partial x_1} f_1 + \frac{\partial h}{\partial x_2} f_2 + \dots + \frac{\partial h}{\partial x_n} f_n$$

It is shown in Ref. 11 that under the assumptions of Theorem 1, we have a desired transformation $(T_1, T_2, \dots, T_n, T_{n+1}, \dots, T_{n+m})$ if and only if we can solve the following partial differential equations:

$$\langle dT_1, (\text{ad}^j f, g_i) \rangle = 0, j = 0, 1, \dots, \kappa_i - 2$$

$$\text{and } i = 1, 2, \dots, m$$

$$\langle dT_{\sigma_1+1}, (\text{ad}^j f, g_i) \rangle = 0, j = 0, 1, \dots, \kappa_i - 2$$

$$\text{and } i = 1, 2, \dots, m$$

...

...

$$\langle dT_{\sigma_m+1}, (\text{ad}^j f, g_i) \rangle = 0, j = 0, 1, \dots, \kappa_i - 2 \quad (4)$$

$$\text{and } i = 1, 2, \dots, m$$

$$\langle dT_{\sigma_1}, f \rangle = \sum_{i=1}^m u_i \langle dT_1, (\text{ad}^{\kappa_i-1} f, g_i) \rangle = T_{n+1}$$

$$\langle dT_{\sigma_2}, f \rangle = \sum_{i=1}^m u_i \langle dT_{\sigma_1+1}, (\text{ad}^{\kappa_i-1} f, g_i) \rangle = T_{n+2}$$

$$\langle dT_n, f \rangle = \sum_{i=1}^m u_i \langle dT_{\sigma_m+1}, (\text{ad}^{\kappa_i-1} f, g_i) \rangle = T_{n+m},$$

where the plus sign is for κ_i odd and the minus sign is for κ_i even, $i = 1, 2, \dots, m$. We also want the matrix

$$\begin{bmatrix} \langle dT_1, (\text{ad}^{\kappa_1-1} f, g_1) \rangle \dots \langle dT_1, (\text{ad}^{\kappa_1-1} f, g_m) \rangle \\ \langle dT_{\sigma_1+1}, (\text{ad}^{\kappa_1-1} f, g_1) \rangle \dots \langle dT_{\sigma_1+1}, (\text{ad}^{\kappa_1-1} f, g_m) \rangle \\ \vdots \\ \langle dT_{\sigma_m+1}, (\text{ad}^{\kappa_m-1} f, g_1) \rangle \dots \langle dT_{\sigma_m+1}, (\text{ad}^{\kappa_m-1} f, g_m) \rangle \end{bmatrix} \quad (5)$$

to be nonsingular. The other coordinate functions of the transformation are found by differentiation: $T_2 = \dot{T}_1 = \frac{\partial T_1}{\partial t}, T_3 = \dot{T}_2, \dots, T_{\sigma_1-2} = \dot{T}_{\sigma_1-1}$, etc.

Note that the last m equations in (4) hold, once we find $T_1, T_{\sigma_1-1}, \dots, T_{\sigma_m-1}$ satisfying the first sets of equations. These last m equations allow us to solve for u_1, u_2, \dots, u_m in terms of $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ (these are the same as

v_1, v_2, \dots, v_m . This fact is extremely important in practice.

To find solutions $T_1, T_{\sigma_1+1}, \dots, T_{\sigma_m-1+1}$ of (4) we move from partial differential equations to ordinary differential equations, as in Ref. 11. We order the n vector fields in C as follows. Let

$$X_1 = (\text{ad}^{\kappa_1-1} f, g_1)$$

$$X_2 = \begin{cases} (\text{ad}^{\kappa_1-1} f, g_2) & \text{if } (\text{ad}^{\kappa_1-1} f, g_2) \in C \\ (\text{ad}^{\kappa_1-2} f, g_1) & \text{if } (\text{ad}^{\kappa_1-1} f, g_2) \notin C \end{cases}$$

$$X_3 = \begin{cases} (\text{ad}^{\kappa_1-1} f, g_3) & \text{if } (\text{ad}^{\kappa_1-1} f, g_3) \in C \\ (\text{ad}^{\kappa_1-1} f, g_1) & \text{if } (\text{ad}^{\kappa_1-1} f, g_3) \notin C \text{ and } (\text{ad}^{\kappa_1-1} f, g_2) \in C \\ (\text{ad}^{\kappa_1-2} f, g_2) & \text{if } (\text{ad}^{\kappa_1-1} f, g_3) \notin C \text{ and } (\text{ad}^{\kappa_1-2} f, g_2) \in C \\ (\text{ad}^{\kappa_1-2} f, g_1) & \text{if } (\text{ad}^{\kappa_1-1} f, g_3) \notin C \text{ and } (\text{ad}^{\kappa_1-2} f, g_2) \notin C \end{cases}$$

$$X_n = g_m$$

We then examine a sequence of ordinary differential equations. Real parameters s_1, s_2, \dots, s_n are introduced by solving in order

$$\frac{dx}{ds_1} = X_1 \text{ with } x(0) = 0$$

$$\frac{dx}{ds_2} = X_2 \text{ with } x(s_1, 0) = x(s_1)$$

$$\frac{dx}{ds_3} = X_3 \text{ with } x(s_1, s_2, 0) = x(s_1, s_2)$$

$$\frac{dx}{ds_n} = X_n \text{ with } x(s_1, s_2, \dots, s_{n-1}, 0) = x(s_1, s_2, \dots, s_{n-1})$$

For $i = 1, 2, \dots, m$ we let p_i equal the subscript of that X_k which is equal to $(\text{ad}^{\kappa_i-1} f, g_i)$. Then we let

$$T_1 = s_{p_1} = s_1$$

$$T_{\sigma_1+1} = s_{p_2}$$

$$T_{\sigma_m-1+1} = s_{p_m}$$

By the inverse function theorem we can solve for s_1, s_2, \dots, s_n as functions of x_1, x_2, \dots, x_n [because matrix (5) is nonsingular], and hence we know $T_1, T_{\sigma_1+1}, \dots, T_{\sigma_m-1+1}$. Examples of this process are given in Refs. 10-12.

Henry Ford (a graduate student of the second author) is currently using the MIT MACSYMA program to construct such transformations and is also investigating numerical methods. Complete solutions for block triangular systems are given in Refs. 6 and 7. An approximation technique using the concept of the tangent model is presented in Ref. 8. This construction is useful in the design of an automatic flight controller for a helicopter. The tangent model gives us the linearization of the transformation about a point (without having the actual transformation) modulo constants

IV. ROBUSTNESS

Our Eq. (1) is a mathematical model for some physical plant that we wish to control. Suppose we are interested in the problem of stabilization. If we choose controls to asymptotically stabilize the model (1) about the origin, do these controls, when applied to nearby nonlinear systems, asymptotically stabilize them about their respective equilibrium points? In particular we hope that the plant is one of these nearby systems, and thus we speak of robustness.

In our work¹³ we consider a transformable nonlinear system (and to make the present discussion easier we assume the transformation is applicable on all of \mathbb{R}^n) and use linear feedback, that is, we choose v_1, v_2, \dots, v_m as linear feedback controls, to asymptotically stabilize the linear system (3). The corresponding controls u_1, u_2, \dots, u_m to use on the mathematical model and the plant are found by solving the last m equations in (4). Given any compact set $K \subset \mathbb{R}^n$ (x_1, x_2, \dots, x_n space) there is an open neighborhood \mathcal{N} of the system (1) (with u_1, u_2, \dots, u_m substituted) in the set of vector fields in $\mathcal{C}^1(K)$ such that if $k \in \mathcal{N}$ then a solution of $\dot{x} = k(x)$ starting at any point $x \in K$ must converge to the equilibrium point of $\dot{x} = k(x)$ in K . Here \mathcal{N} is chosen so that all $k \in \mathcal{N}$ have a unique equilibrium in K . Thus, all nearby vector fields are also stabilized; the proof of this result appears in Ref. 13.

Several important facts are used in developing this theory. One is that if linear feedback controls v_1, v_2, \dots, v_m are taken to asymptotically stabilize the linear canonical system (3), then Lyapunov functions $V(y)$ can be constructed in y -space. However, the composition of a V with a transformation T yields a Lyapunov function for the nonlinear system (1) with controls u_1, u_2, \dots, u_m as computed from our equations. These Lyapunov functions are very valuable in assuring us of the above mentioned robustness. Another useful, but not entirely surprising, discovery is that the eigenvalues of (3) with v_1, v_2, \dots, v_m substituted are the same as the eigenvalues of the linearization of (1) about $(0, 0, \dots, 0)$ and with the corresponding u_1, u_2, \dots, u_m .

V. TRANSFORMATIONS, ESTIMATION, AND NONLINEAR ZEROES

Suppose we take the nonlinear system (1) satisfying the hypotheses of Theorem 1. We add an \mathbb{R}^p -valued output $z = h(x)$, which is \mathcal{G}^m . In transforming system (1) to Brunovsky canonical form (3), we can also transform the output to obtain $z = h(x(y)) = h(x(T))$ for the linear system. We read the output $h(x)$ of the nonlinear system and transform to the output for the linear system. If the states in y -space (where the dynamics are linear) can be estimated, then by the inverse transformation we have estimates of the states in x -space.

Bach¹⁴ has used the concept of transformation for construction of an estimator of aircraft motions.

Continuing to consider systems with outputs, we turn to a discussion of "nonlinear zeroes," or, more precisely, nonlinear zero distributions. Given a linear system with linear output there are various definitions of zeroes, available. We refer the interested reader to Refs. 15-17.

In a recent paper, Isidori et al.¹⁸ discuss nonlinear decoupling via feedback for nonlinear systems. They introduce the concept of (f, g) invariant distribution, which is a nonlinear generalization of the (A, B) invariant subspace of Wonham and Morse¹⁹ and of Basile and Marro.²⁰ Krener and Isidori²¹ give the definition of nonlinear zero distribution, analogous to the ideas of zeroes for linear systems. Two cases are considered: (1) the number of inputs $m \leq p$ = the number of outputs and (2) $m > p$ (here the concept of (h, f) invariant distributions must be introduced). Throughout our discussion we examine only the first case.

Our system is (1) which we now write as

$$\dot{x} = f(x) + g(x)u \quad (6)$$

with \mathcal{G}^m output $z = h(x)$. Here $g(x)$ is the $n \times m$ matrix (g_1, g_2, \dots, g_m) and u the column vector (u_1, u_2, \dots, u_m) and the meaning of $g(x)u$ is obvious.

A distribution Δ on \mathbb{R}^n is an assignment $\Delta(x)$ of a linear subspace of \mathbb{R}^n at each point of \mathbb{R}^n . We assume that Δ is of positive constant dimension and often identify Δ with the set of vector fields in it. We also take Δ to be involutive and regular (from Ref. 18 this means that the quotient set \mathbb{R}^n/Δ is a \mathcal{G}^m manifold). The distribution Δ is (f, g) invariant if there exist $\alpha(x)$ and $\beta(x)$ so that

$$[\tilde{f}, \Delta] \subset \Delta$$

$$[\tilde{g}, \Delta] \subset \Delta$$

where $\tilde{f} = f + g\alpha$ and $\tilde{g} = g\beta$. We say that Δ is a null observable distribution if it is (f, g) invariant and contained in the kernel of $dh = (dh_1, dh_2, \dots, dh_p)$. We suppose that a maximal such distribution exists. From Ref. 21 we deduce that there are initial conditions (not including points where f, g_1, \dots, g_m are linearly dependent) in each level set of the output (the kernel of dh) and controls that force the output to stay in the level set in which it begins. This constant output is the nonlinear generalization of the output being zero for a linear system.

How does the above concept relate to the transformation theory of this paper? Suppose our system (1) [or (6)] is transformable as provided in Theorem 1. Does there exist an m -dimensional \mathcal{G}^m output for which the system is observable for every input, but for which there is not a null observable distribution? The answer is yes if we take $h_1 = T_1, h_2 = T_{\sigma_1+1}, \dots, h_m = T_{\sigma_m+1}$. Since $T_1, T_{\sigma_1+1}, \dots, T_{\sigma_m+1}$ satisfy the equations in (4), no nontrivial (f, g) invariant distribution can be found in the kernel of dh , and the output is observable for every input. This follows from the method in Ref. 18 of building the maximal (f, g) invariant distribution contained in the kernel of dh by differentiating the output repeatedly with respect to time. Let us examine this for the case $m = 1$ and $h(x) = T_1(x)$. To hold the output T_1 constant we must have

$$T_1 = \text{constant}$$

$$T_2 = \dot{T}_1 = \langle dT_1, f \rangle + u \langle dT_1, g_1 \rangle = 0$$

$$T_3 = \dot{T}_2 = \langle dT_2, f \rangle + u \langle dT_2, g_1 \rangle \\ = \langle dT_2, f \rangle + u \langle dT_1, [f, g_1] \rangle = 0$$

(7)

$$T_n = \dot{T}_{n-1} = \langle dT_{n-1}, f \rangle + u \langle dT_{n-1}, g_1 \rangle \\ = \langle dT_{n-1}, f \rangle + u \langle dT_1, (\text{ad}^{n-2} f, g_1) \rangle = 0$$

$$T_{n+1} = \dot{T}_n = \langle dT_n, f \rangle + u \langle dT_n, g_1 \rangle \\ = \langle dT_n, f \rangle + u \langle dT_1, (\text{ad}^{n-1} f, g_1) \rangle = 0$$

The equations $\langle dT_2, g_1 \rangle = -\langle dT_1, [f, g_1] \rangle, \dots, \langle dT_n, g_1 \rangle = -\langle dT_1, (ad^{k_1-1} f, g_1) \rangle$ follow from a Leibniz formula used in developing Eqs. (4) in Ref. 11. Now $\langle dT_1, (ad^j f, g_1) \rangle = 0$ for $j = 0, 1, 2, \dots, k_1-2$ by Eqs. (4) and $\langle dT_1, (ad^{k_1-1} f, g_1) \rangle \neq 0$ from (5). Hence $u = 0$ and from (7)

$$\langle dT_1, f \rangle = \langle dT_1, g_1 \rangle = 0$$

$$\langle dT_2, f \rangle = \langle dT_2, g_1 \rangle = 0$$

$$\langle dT_{n-1}, f \rangle = \langle dT_{n-1}, g_1 \rangle = 0$$

Since $dT_1, dT_2, \dots, dT_{n-1}$ are linearly independent, the initial conditions for which we can apply 0 control and stay in a level set of the output coincides with the set of points where f and g are linearly dependent. Thus, no (f, g) invariant distribution contained in the kernel of dh exists. Also notice that since u does not appear in T_1, T_2, \dots, T_n and dT_1, dT_2, \dots, dT_n are linearly independent, this system is observable for every input.

The main purpose of the above discussion is to show the relationship between the equations for $dT_1, dT_2, \dots, dT_{m-1}$ in (4) and the method of computing null observable distributions in Refs. 18 and 21.

VI. CONCLUSION

We have presented a summary of the theory of controller design via transformations of nonlinear systems to linear systems. The related topics of the construction of transformations, robustness, estimation, and "nonlinear zeroes" have also been treated. As mentioned, a paper discussing the application of our theory is contained elsewhere in these conference proceedings.

REFERENCES

1. P. Brunovsky, A Classification of Linear Controllable System, *Kibernetika (Praha)*, vol. 6, 1970, pp. 173-187.
2. A. J. Krener, On the Equivalence of Control Systems and the Linearization of Nonlinear Systems, *SIAM J. Control*, vol. 11, 1973, pp. 670-676.
3. R. W. Brockett, Feedback Invariants for Nonlinear Systems, IFAC Congress, Helsinki, 1978.
4. B. Jakubczyk and W. Respondek, On Linearization of Control Systems, *Bull. Acad. Pol. Sci., Ser. Sci. Math. Astronom. Phys.*, vol. 28, 1980, pp. 517-522.
5. R. Hermann, The Theory of Equivalence of Pfaffian Systems and Input Systems under Feedback, preprint.
6. G. Meyer and L. Cicolani, A Formal Structure for Advanced Automatic Flight Control Systems, NASA TN-7940, 1975.
7. G. Meyer and L. Cicolani, Applications of Nonlinear System Inverses to Automatic Flight Control Design - System Concepts and Flight Evaluations, AGARDograph on Theory and Application of Optimal Control in Aerospace Systems, P. Kant, ed., 1980.
8. G. Meyer, The Design of Exact Nonlinear Model Followers, 1981 Joint Automatic Control Conference, FA-3A.
9. R. Su, On the Linear Equivalents of Nonlinear Systems, *Systems and Control Letters*, to appear.
10. L. R. Hunt, R. Su, and G. Meyer, Global Transformations of Nonlinear Systems, *IEEE Trans. Autom. Contr.*, to appear.
11. L. R. Hunt, R. Su, and G. Meyer, Design for Multi-Input Nonlinear Systems, Conference on Differential Geometric Control Theory, to appear.
12. L. R. Hunt and R. Su, Control of Nonlinear Time-Varying Systems, 20th IEEE Conference on Decision and Control, San Diego, CA, 1981, pp. 558-563.
13. R. Su, G. Meyer, and L. R. Hunt, Robustness in Nonlinear Control, Conference on Differential Geometric Control Theory, to appear.
14. R. E. Bach, Jr., A Mathematical Model for Efficient Estimation of Aircraft Motions, 6th IFAC Symposium on Identification and System Parameter Estimation, to appear.
15. A.G.J. MacFarlane and N. Karcenas, Poles and Zeroes of Linear Multivariable System: A Survey of the Algebraic, Geometric, and Complex Variable Theory, *Int. J. Contr.*, vol. 24, 1976, pp. 33-74.
16. C. A. Desoer and J. D. Schulman, Zeroes and Poles of Matrix Transfer Functions and Their Dynamical Interpretation, *IEEE Trans. Circuits Systems*, vol. 21, 1974, pp. 3-8.
17. B. F. Wyman and M. K. Sain, The Zero Module and Essential Inverse Systems, *IEEE Trans. Circuits Systems*, vol. 28, 1981, pp. 112-126.
18. A. Isidori, A. J. Krener, C. Gori-Giorgi, and S. Monaco, Nonlinear Decoupling Via Feedback: A Differential Geometric Approach, *IEEE Trans. Autom. Contr.*, vol. 26, 1981, pp. 331-345.
19. W. M. Wonham and A. S. Morse, Decoupling and Pole Assignment in Linear Multivariable Systems: A Geometric Approach, *SIAM J. Control*, vol. 8, 1970, pp. 1-18.
20. G. Basile and G. Marro, Controlled and Conditioned Invariant Subspaces in Linear System Theory, *J. Optimization Theory Appl.*, vol. 3, 1969, pp. 306-315.
21. A. J. Krener and A. Isidori, Nonlinear Zero Distributions, 19th IEEE Conference on Decision and Control, Albuquerque, N.M., 1980, pp. 665-668.

APPLICATIONS TO AERONAUTICS OF THE THEORY OF TRANSFORMATIONS OF NONLINEAR SYSTEMS

George Meyer,* Renjeng Su,** and L. R. Hunt†

NASA Ames Research Center, Moffett Field, California, U.S.A.

ABSTRACT

We discuss the development of a theory, its application to the control design of nonlinear systems, and results concerning the use of this design technique for automatic flight control of aircraft. The theory examines the transformation of nonlinear systems to linear systems. We show how to apply this in practice, in particular, the tracking of linear models by nonlinear plants. Results of manned simulation are also presented.

INTRODUCTION

Suppose we model a physical plant by a nonlinear system

$$\dot{x}(t) = f[x(t)] + \sum_{i=1}^m u_i(t) g_i[x(t)] \quad (1)$$

where f, g_1, \dots, g_m are C^∞ vector fields on \mathbb{R}^n and $f(0) = 0$. If we are to have the output of this plant follow a particular path, then we have a difficult problem to consider. However, if there are new state space coordinates and new controls under which equation (1) becomes a linear system, then our task appears to be much easier because of the known results for controller design on linear systems.

We feel that the following problems are thus of interest:

- Find necessary and sufficient conditions for the system (1) to be transformable to a controllable linear system.
- Show how to use these transformations so that the controller design for nonlinear systems can be reduced to that of linear systems.
- Apply the above theory to the field of aeronautics.

In the next three sections of this paper we discuss the solutions of these problems.

TRANSFORMATION THEORY

The classification of those nonlinear systems that can be transformed to linear systems is actually a subproblem of a much deeper result, the construction of canonical forms for nonlinear systems. We are presently developing a theory for such canonical forms, and in the case that a nonlinear system is transformable as in this paper, the canonical form is actually the Brunovsky (ref. 1) form for a linear system.

Here we concentrate on the transformation theory developed in references 2, 3 and 4. Other significant research in this area is due to Krener (ref. 5), Brockett (ref. 6), Jakubczyk and Respondek (ref. 7), and Hermann (The Theory of Equivalence of Pfaffian Systems and Input Systems under Feedback). We also refer to the early work of the first author in references 8 and 9.

If we are to map our nonlinear system (1) to a controllable linear system, we may as well assume that this linear system is in Brunovsky (ref. 1) canonical form with Kronecker indices $\kappa_1, \kappa_2, \dots, \kappa_m$ satisfying $\sum_{i=1}^m \kappa_i = n$ and $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_m$. Hence this system is

$$\dot{y} = Ay + Bw \quad (2)$$

where A is $n \times n$, B is $n \times m$, $\bar{w} = (w_1, w_2, \dots, w_m)$ are the new controls, A is equal to

*Research Scientist at NASA Ames Research Center.

**Research Associate of National Research Council.

†Research supported by NASA Ames Research Center under the IPA Program and the Joint Services Electronics Program at Texas Tech University, Lubbock, TX and under ONR Contract N00014-77-0-1000.

This paper is declared a work of the U.S. Government and therefore is in the public domain.

AD-A137 619

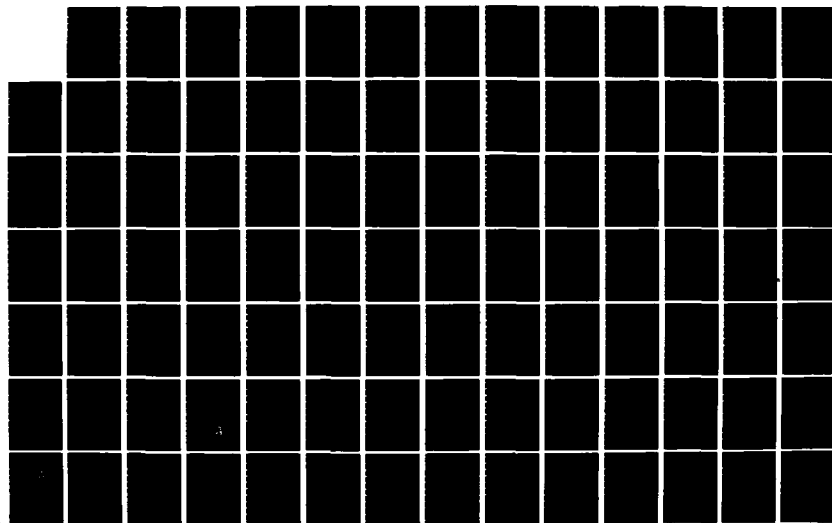
ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICES
ELECTRONICS PROGRAM VO. (U) TEXAS TECH UNIV LUBBOCK
INST FOR ELECTRONIC SCIENCE R SAEKS ET AL. DEC 82
N00014-76-C-1136

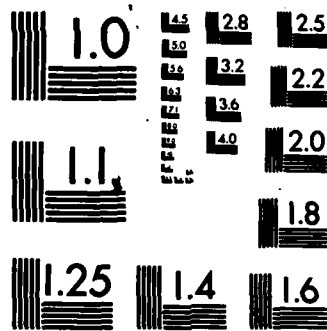
2/4

UNCLASSIFIED

F/G 9/3

NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

We discuss the allowable transformations mapping system (1) to system (2). We want a C^∞ map $Y = (y_1, y_2, \dots, y_n, w_1, w_2, \dots, w_m)$ mapping $R^n \times R^m [(x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m) \text{ space}]$ to $R^n \times R^m [(y_1, y_2, \dots, y_n, w_1, w_2, \dots, w_m) \text{ space}]$ that satisfies the following conditions:

1. Y maps the origin to the origin,
2. y_1, y_2, \dots, y_n are functions of x_1, x_2, \dots, x_n only and have a nonsingular Jacobian matrix,
3. w_1, w_2, \dots, w_m are functions of $x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m$ and for fixed x_1, x_2, \dots, x_n , the $m \times m$ Jacobian matrix of w_1, w_2, \dots, w_m with respect to u_1, u_2, \dots, u_m is nonsingular,
4. Y maps system (1) to system (2),
5. Y is a one-to-one map of $R^n \times R^m$ onto $R^n \times R^m$.

Next we introduce some basic definitions from differential geometry.

If f and g are C^∞ vector fields on R^n , the Lie bracket of f and g is

$$[f, g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g$$

where $\frac{\partial g}{\partial x}$ and $\frac{\partial f}{\partial x}$ are Jacobian matrices. We let

$$(\text{ad}^0 f, g) = g$$

$$(\text{ad}^1 f, g) = [f, g]$$

$$(\text{ad}^2 f, g) = [f, (f, g)]$$

$$(\text{ad}^k f, g) = [f, (\text{ad}^{k-1} f, g)]$$

A collection of C^∞ vector fields h_1, h_2, \dots, h_r is involutive if there exists C^∞ functions γ_{ijk} such that

$$[h_i, h_j](x) = \sum_{k=1}^r \gamma_{ijk}(x) h_k(x), \quad 1 \leq i, j \leq r, i \neq j.$$

Let $\langle \dots \rangle$ denote the duality between one forms and vector fields. If $\omega = \omega_1 dx_1 + \omega_2 dx_2 + \dots + \omega_n dx_n$ is a differentiable one form and f a vector field on R^n , then

$$\langle \omega, f \rangle = \omega_1 f_1 + \omega_2 f_2 + \dots + \omega_n f_n$$

To state the main result from reference 4 giving necessary and sufficient conditions for transforming system (1) to system (2) we need the following sets:

$$C = \{g_1, [f, g_1], \dots, (\text{ad}^{k_1-1} f, g_1), g_2, [f, g_2], \dots, (\text{ad}^{k_2-1} f, g_2), \dots, g_m, [f, g_m], \dots, (\text{ad}^{k_m-1} f, g_m)\}$$

$$C_j = \{g_1, [f, g_1], \dots, (\text{ad}^{k_j-2} f, g_1), g_2, [f, g_2], \dots, (\text{ad}^{k_j-2} f, g_2), \dots, g_m, [f, g_m], \dots, (\text{ad}^{k_j-2} f, g_m)\} \text{ for } j=1, 2, \dots, m.$$

Theorem 2.1 There exists a transformation $Y = (y_1, y_2, \dots, y_n, w_1, w_2, \dots, w_m)$ satisfying conditions 1) through v) above if and only if on R^n

- 1) the set C spans an n dimensional space,
- 2) each set C_j is involutive for $j = 1, 2, \dots, m$, and,
- 3) the span of C_j equals the span of $C_j \cap C$ for $j = 1, 2, \dots, m$.

Let $\sigma_1 = k_1, \sigma_2 = k_1 + k_2, \dots, \sigma_m = k_1 + k_2 + \dots + k_m = n$. Then the transformation is constructed in reference 4 by solving the partial differential equations

$$\begin{aligned} \langle dy_i, (\text{ad}^j f, g_i) \rangle &= 0, j=0, 1, \dots, k_i-2 \text{ and } i=1, 2, \dots, m, \\ \langle dy_{\sigma_{j+1}}, (\text{ad}^j f, g_i) \rangle &= 0, j=0, 1, \dots, k_i-2 \text{ and } i=1, 2, \dots, m, \end{aligned}$$

$$\langle dy_{\sigma_{m-1}+1}, (ad^j f, g_1) \rangle = 0, j=0,1,\dots, \kappa_m-2 \text{ and } i=1,2,\dots,m. \quad (3)$$

$$\langle dy_{\sigma_1}, f \rangle + \sum_{i=1}^m u_i \langle dy_{\sigma_1}, g_i \rangle = w_1$$

$$\langle dy_{\sigma_2}, f \rangle + \sum_{i=1}^m u_i \langle dy_{\sigma_2}, g_i \rangle = w_2$$

$$\langle dy_n, f \rangle + \sum_{i=1}^m u_i \langle dy_n, g_i \rangle = w_m$$

where the matrix

$$\begin{bmatrix} \langle dy_1, (ad^{\kappa_1-1} f, g_1) \rangle & \dots & \langle dy_1, (ad^{\kappa_1-1} f, g_m) \rangle \\ \langle dy_{\sigma_1+1}, (ad^{\kappa_2-1} f, g_1) \rangle & \dots & \langle dy_{\sigma_1+1}, (ad^{\kappa_2-1} f, g_m) \rangle \\ \vdots & & \vdots \\ \langle dy_{\sigma_{m-1}+1}, (ad^{\kappa_m-1} f, g_1) \rangle & \dots & \langle dy_{\sigma_{m-1}+1}, (ad^{\kappa_m-1} f, g_m) \rangle \end{bmatrix} \quad (4)$$

is nonsingular.

It can be shown that matrix (4) being invertible means we can solve for u_1, u_2, \dots, u_m in terms of w_1, w_2, \dots, w_m in the last m equations in equation (3).

Equation (3) can be formally solved by considering a sequence of ordinary differential equations as in reference 4, but we shall not mention details here.

If a nonlinear system is transformable to a linear system, we study the process of using the transformation to construct a controller for the nonlinear system.

TRANSFORMATIONS IN CONTROLLER DESIGN

Let $Y = (y_1, y_2, \dots, y_n, w_1, w_2, \dots, w_m)$ be the transformation from system (1) to system (2) as before. The structure of the control system using transformation theory is illustrated in figure 1. The design scheme is implemented on the "linear part" of the diagram, and this system is in Brunovsky form.

We ask that the output of the nonlinear system follow a particular path which corresponds to a trajectory for the output of the linear model. If we know how to design for the linear system, then we actually have a tracking of a linear model by a nonlinear plant.

Linear design is used to generate an open loop command w_c for the system (2), and we find the corresponding y coordinates y_c by plugging w_c into equation (2). The transformation Y maps the measured x space to y space and y is compared to y_c and the difference is an error e_y . The regulator yields a control δw which sends e_y to zero, and variations in plant dynamics and disturbances are compensated for in this way.

The controls w_c and δw are added and transformed through the inverse map R (actually $w_c + \delta w$ is substituted into the last m equations in equation (3) and $u = (u_1, u_2, \dots, u_m)$ is generated) to obtain a control which is applied to the plant. Thus we have an exact model follower, and the difficult problem of finding an open loop control and the regulator control are constrained to the linear system.

The remainder of this paper contains the application of the transformation theory to aeronautics.

AUTOMATIC FLIGHT CONTROLLER DESIGN

The aircraft will be represented by a rigid body moving in 3-dimensional space in response to gravity, aerodynamics and propulsion.

The state

$$x = \begin{pmatrix} r \\ v \\ C \\ \omega \end{pmatrix} \in X \subset R^3 \times R^3 \times SO(3) \times R^3 \quad (5)$$

where r and v are inertial coordinates of body center of mass position and velocity, respectively; C is the direction cosine matrix of body fixed axes relative to the runway fixed axes (inertial), and ω is the angular velocity.

The controls,

$$u = \begin{pmatrix} u^M \\ u^P \end{pmatrix} \in U \subset R^3 \times R \quad (6)$$

where u^M is the 3-axis moment control such as ailerons, elevator and rudder in a conventional aircraft or roll cyclic, pitch cyclic and tail rotor collective in a helicopter; and u^P controls power - throttle in a conventional aircraft, and the main rotor collective in a helicopter. The state equation consists of the translational and rotational kinematic and dynamic equations:

$$\begin{aligned} \dot{r} &= v \\ \dot{v} &= f^F(x, u) \\ \dot{C} &= S(\omega)C \\ \dot{\omega} &= f^M(x, u) \end{aligned} \quad (7)$$

where f^F and f^M are the total force and moment generation processes and $x \in X$. We wish to transform equation (7) into a linear system.

In general, f^M is invertible with respect to the (vector) pair $(\dot{\omega}, u^M)$, and, for the specific class of helicopter maneuvers being considered (i.e., no 360° rolls), f^F is invertible with respect to the (scalar) pair (\dot{v}_3, u^P) . Thus, a function $h: X \times R^h \rightarrow U$ can be constructed such that if

$$\begin{pmatrix} u^M \\ u^P \end{pmatrix} = h(r, v, C, \omega, \dot{\omega}_0, \dot{v}_{30}) \quad (8)$$

then

$$\begin{aligned} \dot{\omega} &= \dot{\omega}_0 \\ \dot{v}_3 &= \dot{v}_{30} \end{aligned} \quad (9)$$

for all admissible maneuvers. That is, angular and vertical accelerations can be chosen as the new set of independent controls in which case the state equation may be written as follows

$$\begin{aligned} \dot{r} &= v \\ \dot{v} &= f^0(r, v, C, \dot{v}_{30}) + cf^1(r, v, C, \dot{v}_{30}, \omega, \dot{\omega}_0) \\ \dot{C} &= s(\omega)C \\ \dot{\omega} &= \dot{\omega}_0 \end{aligned} \quad (10)$$

where $c = 1$, and $f^1(r, v, C, 0, 0, 0) = 0$ for all admissible maneuvers.

The function f^0 is invertible with respect to $((\dot{v}_1, \dot{v}_2, E_3(\psi)), C)$ where $E_3(\psi)$ is an elementary rotation about the z-axis and represents the heading of the helicopter. Thus, a function $h^f: R^6 \times SO(2) \rightarrow SO(3)$ can be constructed such that if

$$C_0 = h^f(r, v, \dot{v}_0, E_3(\psi_0)) \quad (11)$$

then

$$\dot{v} = \dot{v}_0 \quad (12)$$

Equations (8) and (11) are the trim equations of the process equation (10) (with $c = 0$). That is, for a given path $(r(t), E_3(\psi(t)))$, $t \geq 0$ with $\dot{v}_3(t) = 0$, the corresponding state and control may be constructed as follows

$$\begin{aligned}
r_0 &= r(t) \\
v_0 &= \dot{r}(t) \\
C_0 &= n^2(r_0, v_0, \dot{v}(t), E_3, (v(t))) \\
\omega_0 &= q(\dot{C}C^c) \\
\dot{\omega}_0 &= (\omega_0)^{\cdot} \\
u_0 &= h(r_0, v_0, C_0, \omega_0, \dot{\omega}_0, 0)
\end{aligned} \tag{13}$$

where the function q extracts u from $\dot{C}C^c = S(u)$. The required time derivatives in equation (13) can be computed provided that the path (r, E_3) is generated by the system diagrammed in figure 2 where \circ represents a scalar integrator and y_0 the control (v in the previous section is y^5 here).

We construct an approximation to the linearizing transformation as follows: Y_1, R_1, Q are constructed so that

$$\begin{aligned}
y &= Y(x) \approx Y_0(x_0) + Y_1 \delta x = y_0 + Y_1 \delta x \\
u &= R(x, y^5) \approx u_0 + R_1 \delta y^5 + Q \delta x
\end{aligned} \tag{14}$$

Here Y_0 is the transformation when $c = 0$, and δx (and δy^5) is the perturbation about the nominal x_0 (and y_0^5) given in equation (13) (and figure 2).

From equation (10) with $c = 0$, it follows that $(C = (I + S(c)C_0)$

$$\begin{aligned}
(\delta r)^{\cdot} &= \delta v \\
(\delta v)^{\cdot} &= \frac{\partial f}{\partial r} \delta r + \frac{\partial f}{\partial v} \delta v + \frac{\partial f}{\partial c} c + \frac{\partial f}{\partial v_3} \dot{v}_3 \\
(c)^{\cdot} &= \delta u \\
(\delta u)^{\cdot} &= \delta \dot{\omega}_0
\end{aligned} \tag{15}$$

where c is attitude perturbation.

The pattern of equation (15) after some rearrangement of coordinates is shown in equation (16).

$$\begin{bmatrix} r_1 \\ r_2 \\ v_1 \\ v_2 \\ c_1 \\ c_2 \\ c_3 \\ r_3 \\ u_1 \\ u_2 \\ u_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0 & I & 0 & 0 \\ 0 & C_1 & C_2 & C_3 & 0 & C_4 \\ 0 & 0 & 0 & I \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ v_1 \\ v_2 \\ c_1 \\ c_2 \\ c_3 \\ r_1 \\ u_1 \\ u_2 \\ u_3 \\ v_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 & C_5 \\ 0 \\ I \end{bmatrix} \begin{bmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \dot{u}_3 \\ \dot{v}_3 \end{bmatrix} \tag{16}$$

In the present case of the helicopter, C_1, C_4 , and C_5 are negligible. Their effect will be controlled by the regulator.

The transformations

$$\begin{bmatrix} \delta y_1^1 \\ \delta y_2^1 \\ \delta y_1^2 \\ \delta y_2^2 \\ \delta y_1^3 \\ \delta y_2^3 \\ \delta y_3^3 \\ \delta y_4^3 \\ \delta y_1^4 \\ \delta y_2^4 \\ \delta y_3^4 \\ \delta y_4^4 \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & C_2 & C_3 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & C_2 & C_3 \\ 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} \delta r_1 \\ \delta r_2 \\ \delta r_1 \\ \delta r_2 \\ e_1 \\ e_2 \\ e_3 \\ \delta r_3 \\ \delta \omega_1 \\ \delta \omega_2 \\ \delta \omega_3 \\ \delta r_3 \end{bmatrix} \quad (17)$$

$$\begin{bmatrix} \delta \dot{\omega}_1 \\ \delta \dot{\omega}_2 \\ \delta \dot{\omega}_3 \\ \dot{v}_3 \end{bmatrix} = \begin{bmatrix} C_2^{-1} & -C_2^{-1}C_3 \\ 0 & I \end{bmatrix} \begin{bmatrix} \delta y_1^5 \\ \delta y_2^5 \\ \delta y_3^5 \\ \delta y_4^5 \end{bmatrix} \quad (18)$$

take the system in equation (16) (with $C_1, C_4, C_5 = 0$) into the canonic system:

$$(\delta y)^* = \begin{bmatrix} 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ 0 & 0 & 0 & 0 \end{bmatrix} \delta y + \begin{bmatrix} 0 \\ 0 \\ 0 \\ I \end{bmatrix} \delta y^5 \quad (19)$$

Thus, the linearizing transformation (Y and R in figure 1) is constructed.

That the accuracy of the transformation is adequate may be seen from the results of the simulation of the flight experiment to be briefly summarized next.

The test consists in automatically flying a trajectory which exercises the system over a wide range of flight conditions as shown in figures 4 and 5.

Thus, the test takes the helicopter from hover (WP1) to high speed (150 ft/sec) turning acceleration, ascending flight.

Figure 6 shows the resulting tracking errors.

As can be seen, position tracking error e_r is quite small. The acceleration errors e_a , which is due to the neglected terms in the construction of the linearizing transformation is also quite small. In summary, the resulting performance of the system is good.

REFERENCES

1. Brunovsky, P.: A Classification of Linear Controllable Systems. Kibernetika (Prague). Vol. 6, 1970, pp. 173-188.
2. Su, R.: On the Linear Equivalents of Nonlinear Systems, Systems and Control Letters. Vol. 2, No. 1. (To appear in print 1982.)
3. Hunt, L. R.; Su, R.; and Meyer, G.: Global Transformations of Nonlinear Systems. IEEE Trans. on Automatic Control. Vol. 27, No. 6. (To appear in print 1982.)
4. Hunt, L. R.; Su, R.; and Meyer, G.: Multi-input Nonlinear Systems. Differential Geometric Control Theory Conference. Birkhauser Boston, Cambridge. (To appear in print 1982.)
5. Krener, A. J.: On the Equivalence of Control Systems and Linearization of Nonlinear Systems. SIAM Journal of Control. Vol. 11, 1973, pp. 670-676.
6. Brockett, R. W.: Feedback Invariants for Nonlinear Systems. IFAC Congress, Helsinki, 1978.
7. Jakubczyk, B.; and Respondek, W.: On Linearization of Control Systems. Bull. Acad. Pol. Sci., Ser. Sci., Math. Astronom. Phys., Vol. 28, 1980, pp. 517-522.
8. Meyer, G.; and Cicolani, L.: A Formal Structure for Advanced Flight Control Systems. NASA TN D-7940, 1975.
9. Meyer, G.; and Cicolani, L.: Applications of Nonlinear System Inverses to Automatic Flight Control Design - System Concepts and Flight Evaluations. AGARDograph on Theory and Application of Optimal Control in Aerospace Systems, P. Kant, ed., 1980.

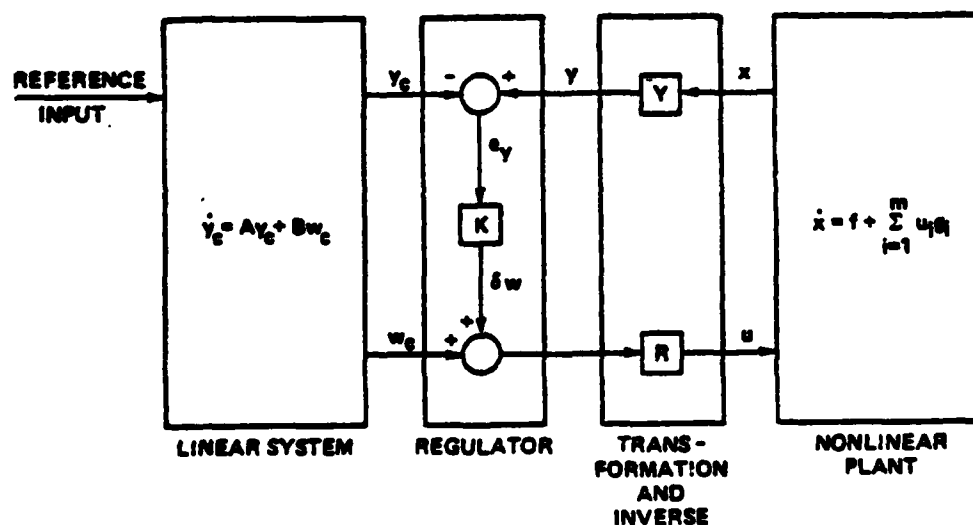


Figure 1. Structure of the Control System

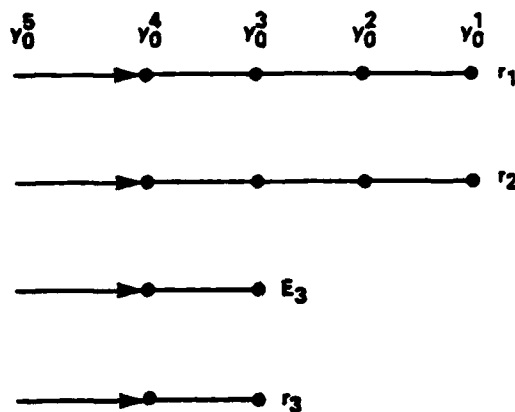


Figure 2. Reference Canonic System

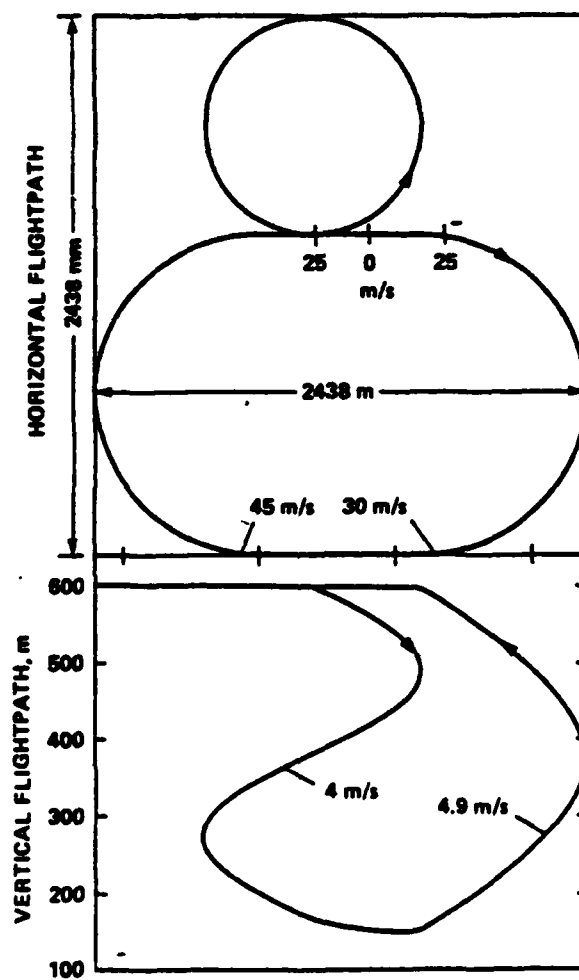


Figure 3. Experimental Flightpath Shown in Horizontal and Vertical Planes

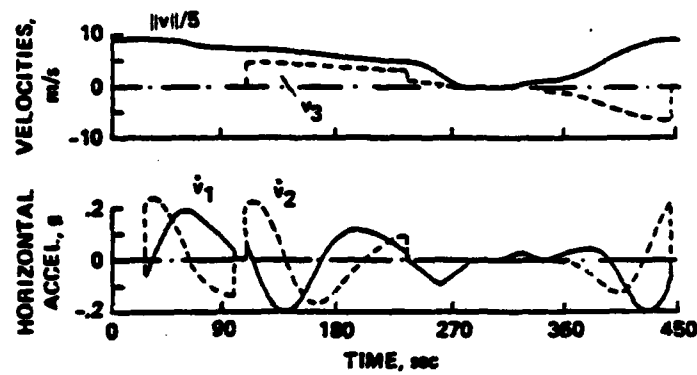


Figure 4. Speed and Acceleration of Experimental Trajectory

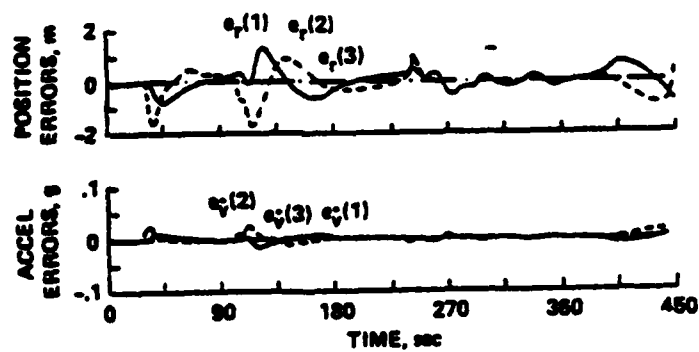


Figure 5. Tracking Errors

DESIGN FOR MULTI-INPUT NONLINEAR SYSTEMS

L. R. Hunt*, Renjeng Su**, and G. Meyer†

Abstract: Consider the multi-input nonlinear system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^m u_i(t)g_i(x(t)) \text{ ,}$$

where f, g_1, \dots, g_m are \mathcal{C}^∞ vector fields on some neighborhood of the origin in \mathbb{R}^n and $f(0) = 0$. We present necessary and sufficient conditions for this system to be transformed to a controllable linear system. Our results are constructive and depend upon the solutions of overdetermined systems of partial differential equations. Moreover, we indicate how this theory is applied to build an automatic flight controller for vertical and short takeoff (VSTOL) aircraft. Flight-test simulation results are presented.

1. Introduction. We examine nonlinear systems of the type

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^m u_i(t)g_i(x(t)) \text{ , } \quad \dot{x} = \frac{dx}{dt} \text{ ,} \quad (1)$$

where f, g_1, \dots, g_m are \mathcal{C}^∞ vector fields defined on some open neighborhood of the origin in \mathbb{R}^n and $f(0) = 0$. Our goal is to find necessary and sufficient conditions on (1) which guarantee that this system can be transformed to a controllable linear system, again for (x_1, x_2, \dots, x_n) in some neighborhood of the origin in \mathbb{R}^n . Of course, we want our transformations to map the origin to the origin, to have a nonsingular Jacobian matrix, and to be one-to-one. For the purpose of applications, it is convenient to have a method to construct a transformation, rather than just conditions under which a transformation exists. Therefore, such a method is introduced.

Meyer and Cicolani in [8] and [9] have considered such transformations for block triangular systems. Krener [7] found conditions if the

transformations are restricted to state space coordinate changes. In the single-input case, Brockett [1] presented assumptions under which a nonlinear system and a controllable linear system are equivalent if coordinate changes and additive state feedback are used. With the addition of input space coordinate changes for a multi-input system, Jakubczyk and Respondek [5] discovered necessary and sufficient conditions for a nonlinear system to be equivalent to a controllable linear system.

Recently, the second author [13] proved a result for transforming a single input nonlinear system ($m = 1$ and $g_1 = g$) to a linear system, where the conditions under which transformations exist are more general than those mentioned above. Essentially, the system (1) is "equivalent (in a neighborhood of the origin)" to a linear system in integrator form if and only if $g, [f, g], (ad^2 f, g), \dots, (ad^{n-1} f, g)$ are linearly independent and $g, [f, g], (ad^2 f, g), \dots, (ad^{n-2} f, g)$ are involutive in some open neighborhood of the origin in \mathbb{R}^n . Here $[\cdot, \cdot]$ denotes the Lie bracket, $(ad^2 f, g) = [f, [f, g]]$, etc.

The purpose of the paper is to generalize the theory in [13] to multi-input systems. First, we establish a notion of equivalence between systems and determine the types of transformations of interest to us. In fact, we derive a system of partial differential equations, the solution of which gives us a transformation. We remark that our sufficient conditions for the existence of a transformation are weaker than those found in [5]. Most importantly, in addition to showing the existence of a transformation, we actually present a constructive proof.

Since we are mapping to controllable linear systems and each such linear system has a Brunovsky [2] canonical form based on its Kronecker indices (see [6] or [16]), we may as well choose a set of Kronecker indices and a canonical form and ask for necessary and sufficient conditions to transform our system (1) to this canonical form. Suppose we take positive integers (the Kronecker indices) $\kappa_1, \kappa_2, \dots, \kappa_m$ with $\kappa_1 + \kappa_2 + \dots + \kappa_m = n$ and $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_m$ and the Brunovsky form associated with these indices. Assume that g_1, g_2, \dots, g_m are linearly independent near the origin (this can be weakened). With a possible reordering of g_1, g_2, \dots, g_m , we can transform our nonlinear system to the linear system in this Brunovsky form if and only if in some neighborhood of the origin in \mathbb{R}^n :

- i) The set $C = \{g_1, [f, g_1], \dots, (\text{ad}^{\kappa_1-1} f, g_1), g_2, [f, g_2], \dots, (\text{ad}^{\kappa_2-1} f, g_2), \dots, g_m, [f, g_m], \dots, (\text{ad}^{\kappa_m-1} f, g_m)\}$ spans an n -dimensional space,
- ii) The sets $C_j = \{g_1, [f, g_1], \dots, (\text{ad}^{\kappa_j-2} f, g_1), g_2, [f, g_2], \dots, (\text{ad}^{\kappa_j-2} f, g_2), \dots, g_m, [f, g_m], \dots, (\text{ad}^{\kappa_j-2} f, g_m)\}$ are involutive for $j = 1, 2, \dots, m$, and
- iii) The span of each C_j is equal to the span of $C_j \cap C$.

We obtain a transformation $T: \mathbb{R}^{n+m} \rightarrow \mathbb{R}^{n+m}$ where $(x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m)$ are the variables in the domain and $(T_1, T_2, \dots, T_{n+m})$ are the variables in the range. Of course x_1, x_2, \dots, x_n are the state space variables and u_1, u_2, \dots, u_m are the controls for our nonlinear system (1). Likewise T_1, T_2, \dots, T_n are the state variables and $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ are the controls for our linear system in canonical form. The existence of such a mapping follows from the Frobenius Theorem, and we give a constructive proof of the transformation.

In [5] the additional assumptions made for the transformation problem are that each of the sets

$$C_j^i = \{g_1, [f, g_1], \dots, (\text{ad}^{\kappa_j-i} f, g_1), g_2, [f, g_2], \dots, (\text{ad}^{\kappa_j-i} f, g_2), \dots, g_m, [f, g_m], \dots, (\text{ad}^{\kappa_j-i} f, g_m)\}$$

are involutive for $j = 1, 2, \dots, m$ and $i = 3, 4, \dots, \kappa_j$. However, it is quite surprising that we can show that these extra conditions are a consequence of our assumptions i), ii), and iii) above. This fact is apparent in the approach we take through the topic of partial differential equations and indicates the power of our technique.

At NASA Ames Research Center, the theory provides a formal mathematical framework for our design technique involving an automatic flight controller for VSTOL aircraft. The highly nonlinear mathematical models are transformed to controllable linear systems, where known techniques can be applied to build a controller. The partial differential equations approach as presented in this paper is the basis for the construction of our transformations. In many cases, transformations can be built and researchers are now working on methods to find transformations by the symbolic computation of the MIT MACSYMA program and by numerical methods. We also have a technique for providing

transformations, which are computed on-line along the system trajectory [10]. Furthermore, we are able to build a transformation by first constructing the exact mapping on a lower dimensional submanifold of R^n and then extrapolating off this submanifold using the partial differential equations.

An application of transformation theory to the design of exact nonlinear model followers is given in Section 2. In Section 3 of this paper we present definitions and preliminaries. Section 4 contains our definitions of equivalence for systems and the partial differential equations we must solve to find a transformation. In Section 5 we give a constructive proof of our result involving transformations and present an example. Also we exhibit a method of choosing the appropriate Kronecker indices given the nonlinear system (1). An Appendix contains the results of flight tests for automatic control of the UH-1H helicopter.

2. Applications.

Before we begin our discussion of conditions under which transformations from nonlinear to linear systems exist and how such transformations can be constructed, we first present an application of the theory.

Suppose we have a nonsingular transformation T mapping our system (1) to a time-invariant and controllable linear system with state variable $y = (y_1, y_2, \dots, y_n)$ and control variable $v = (v_1, v_2, \dots, v_m)$. We control the plant (the nonlinear system) by controlling the linear system. The proposed structure of the complete control system is specified in figure 1. Note that the design is carried out on the "linear side" of the transformation, and in our theory this linear system is in Brunovsky form.

Suppose we wish the output of the linear system to perform a particular task corresponding to a similar task for the nonlinear system in x space. Linear design is used to find an open-loop command v_0 , and we obtain the corresponding y_0 coordinates by substituting v_0 into the linear system. The transformation T maps from x space to y space and y is compared to y_0 to yield an error e_y . The regulator stabilizes out the difference using a control δv ; disturbances and variations in plant dynamics are handled in this

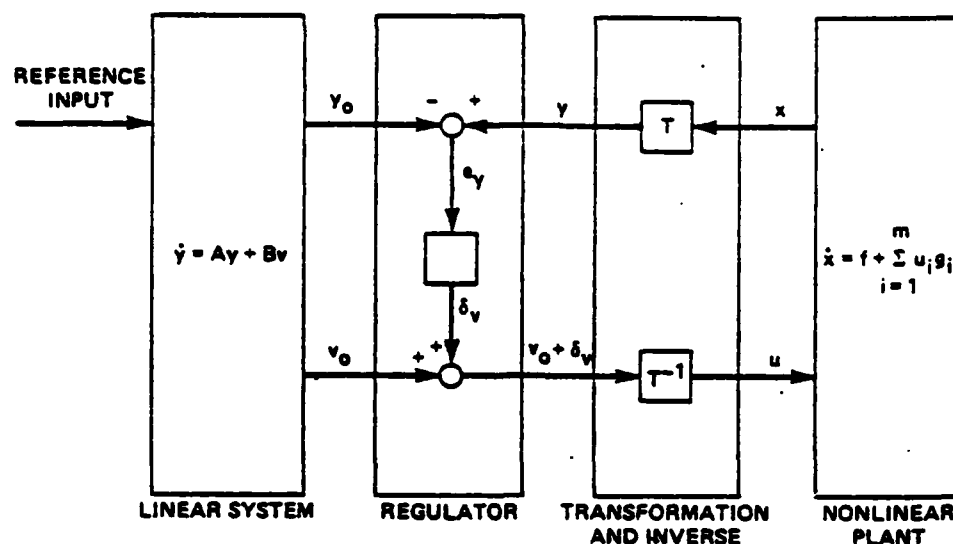


Figure 1

way. The controls v_0 and δv are added and transformed through the inverse map T^{-1} to give a control $u = (u_1, u_2, \dots, u_m)$ which is applied to the plant. By this we generate an exact model follower. The difficult tasks of finding the open-loop control and the regulated control are performed on the linear system that make an easier design possible.

The aerodynamic characteristics and operational requirements of modern aircraft present the control system designer with requirements that are increasingly difficult to solve using the standard control system design methods. There can be strong, multiaxis, highly coupled nonlinearities, and anticipated operational capabilities require that the aircraft be precisely controlled over a substantial portion of the flight envelope that encompasses the range of nonlinear aerodynamics. Thus, the nonlinearity is an essential part of the design problem.

The design approach discussed in this paper and in reference [10] has been applied to several aircraft of increasing complexity. The completely automatic flight-control system was first tested in flight on a DHC-6. This test required the aircraft to operate over a major part of the flight envelope, and the system performed well despite disturbances and plant variations (see [14]). Next, the approach was applied to the Augmentor Wing Jet STOL Research aircraft and successfully flight tested [9]. Pilot inputs were incorporated into the design method in [15]. The approach was also applied to control an A-7 for carrier landing and tested in manned simulation (see [11], [12]). The design method is currently being used for the UH-1H helicopter.

3. Preliminaries.

We now give basic definitions and results from differential geometry and from the theory of controllable linear systems.

For \mathcal{C}^∞ vector fields f and g on \mathbb{R}^n we define the Lie bracket of f and g

$$[f, g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g$$

where $\partial g / \partial x$ and $\partial f / \partial x$ are Jacobian matrices. We can introduce successive Lie brackets $[f, [f, g]]$, $[g, [f, g]]$, etc., and define

$$\begin{aligned} (\text{ad}^0 f, g) &= g \\ (\text{ad}^1 f, g) &= [f, g] \\ (\text{ad}^2 f, g) &= [f, [f, g]] \\ &\vdots \\ (\text{ad}^k f, g) &= [f, (\text{ad}^{k-1} f, g)]. \end{aligned}$$

A set of \mathcal{C}^∞ vector fields $\{f_1, f_2, \dots, f_r\}$ on \mathbb{R}^n is involutive if there exist \mathcal{C}^∞ functions $\gamma_{ijk}(x)$ such that

$$[f_i, f_j](x) = \sum_{k=1}^r \gamma_{ijk}(x) f_k(x), \quad 1 \leq i, j \leq r, i \neq j.$$

If f_1, f_2, \dots, f_r is an involutive collection of linearly independent vector fields and $x_0 \in \mathbb{R}^n$, then by the Frobenius Theorem there exists a unique r dimensional \mathcal{C}^∞ submanifold S of \mathbb{R}^n through x_0 with the tangent space to S at each $x \in S$ being the space spanned by $f_1(x), f_2(x), \dots, f_r(x)$. We say that S is the unique integral manifold of f_1, f_2, \dots, f_r through x_0 .

Suppose f is a \mathcal{C}^∞ vector field on \mathbb{R}^n and h is a \mathcal{C}^∞ function with gradient dh . The Lie derivative of h with respect to f is

$$L_f(h) = \langle dh, f \rangle.$$

where $\langle \cdot, \cdot \rangle$ denotes the duality between one forms and vector fields. If ω is a \mathcal{C}^∞ one form on \mathbb{R}^n , we define the Lie derivative of ω with respect to f

$$L_f(\omega) = \left(\frac{\partial \omega^*}{\partial x} f \right)^* + \omega \frac{\partial f}{\partial x},$$

where * denotes transpose and $\partial \omega^* / \partial x$ and $\partial f / \partial x$ are Jacobian matrices.

The three kinds of Lie differentiations just defined are related by the Leibnitz type formula

$$L_f(\omega, g) = \langle L_f(\omega), g \rangle + \langle \omega, [f, g] \rangle, \quad (2)$$

with f and ω as before and g as a \mathcal{G}^m vector field. For a \mathcal{G}^m function h we also have $dL_f(h) = L_f(dh)$.

We introduce the Kronecker indices and the Brunovsky [2] canonical forms. Consider the linear system

$$\dot{y} = Ay + Bv \quad (3)$$

where y and v are n vectors and m vectors, respectively, and A and B are matrices of the appropriate size. For our discussion we assume that A and B are time invariant, B has rank m , and the span of $\{B, AB, \dots, A^{n-1}B\}$ is n dimensional. Set $r_0 = \text{rank } B$,

$$r_j = \text{rank } \{B, AB, \dots, A^j B\} - \text{rank } \{B, AB, \dots, A^{j-1} B\}, \\ 1 \leq j \leq n-1.$$

Obviously, $0 \leq r_j \leq m$ for $0 \leq j \leq n-1$ and

$$\sum_{j=0}^{n-1} r_j = n.$$

We define the Kronecker indices $\kappa_1, \kappa_2, \dots, \kappa_m$ by κ_i is the number of r_j 's that are $\geq i$ and we remark that $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_m$ and

$$\sum_{j=0}^m \kappa_j = n.$$

Then the system (3) is equivalent to a linear system (in Brunovsky canonical form)

$$\dot{y} = \hat{A}y + \hat{B}v \quad (4)$$

and \hat{B} equals

$$\begin{array}{l} \kappa_1 \left\{ \begin{array}{l} 0 \ 0 \ . \ . \ . \ . \ . \ 0 \\ 0 \ 0 \ . \ . \ . \ . \ . \ 0 \\ \vdots \\ \vdots \\ \vdots \\ 1 \ 0 \ . \ . \ . \ . \ . \ 0 \\ \hline 0 \ 0 \ . \ . \ . \ . \ . \ 0 \\ 0 \ 0 \ . \ . \ . \ . \ . \ 0 \\ \vdots \\ \vdots \\ \vdots \end{array} \right. \\ \kappa_2 \left\{ \begin{array}{l} 0 \ 0 \ . \ . \ . \ . \ . \ 0 \\ 0 \ 0 \ . \ . \ . \ . \ . \ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \ 1 \ 0 \ . \ . \ . \ . \ 0 \\ \hline \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \end{array} \right. \\ \kappa_m \left\{ \begin{array}{l} 0 \ 0 \ . \ . \ . \ . \ . \ 0 \\ 0 \ 0 \ . \ . \ . \ . \ . \ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \ 0 \ . \ . \ . \ . \ . \ 1 \end{array} \right. \end{array}$$

Later, in the coordinate system we use (y_1, y_2, \dots, y_n)
 $= (T_1, T_2, \dots, T_n)$ and $(v_1, v_2, \dots, v_m) = (T_{n+1}, T_{n+2}, \dots, T_{n+m})$.

4. Equivalence of Systems.

The definitions of \mathcal{F} -transformations and \mathcal{F} -related systems for single input systems are given in [13]. We extend these definitions to multi-input systems. Let U be an open neighborhood of the origin in \mathbb{R}^{n+m} space.

Definition 4.1: A \mathcal{F} -transformation T with domain U is a diffeomorphism onto an open neighborhood of the origin in \mathbb{R}^{n+m} which is nonsingular and maps the origin to the origin.

In our theory U is essentially $V \times \mathbb{R}^m$, where V is an open neighborhood of the origin in \mathbb{R}^n , the state space.

Let $(x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m)$ and $(y_1, y_2, \dots, y_n, v_1, v_2, \dots, v_m)$ ($= (T_1, T_2, \dots, T_n, T_{n+1}, T_{n+2}, \dots, T_{n+m})$) denote the state and control variables in U and $T(U)$, respectively. Suppose we examine two systems

$$S_1: \dot{x} = a(x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m)$$

and

$$S_2: \dot{y} = b(y_1, y_2, \dots, y_n, v_1, v_2, \dots, v_m) \text{ with respective state trajectory functions } \phi \text{ and } \psi.$$

Definition 4.2: The system S_1 is \mathcal{F} -related to the system S_2 if there exists a \mathcal{F} -transformation T on U such that for each state $x_0 \in V$ and each admissible control u the following holds. If we let $y_0 = T(x_0, u(0))$ and $T(\phi(t; x_0, u), u(t)) = (y(t), v(t))$ whenever $\phi(t; x_0, u)$ is a state in U , then $y(t) = \psi(t; y_0, v)$.

If S_1 is \mathcal{F} -related to S_2 by the transformation T with domain U we say that S_1 is T -related to S_2 . In fact, by the following statement it makes sense to say that S_1 is T -equivalent to S_2 . In [13] it is shown that the \mathcal{F} -relation for single input systems is actually an equivalence relation, and the proof can be generalized to our present case.

(5) We are particularly interested in mapping (via a \mathcal{F} -transformation) the nonlinear system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^m u_i(t) g_i(x(t)), \quad (1)$$

with $f(0) = 0$, to the controllable linear system (4) in Brunovsky canonical form with indices $\kappa_1, \kappa_2, \dots, \kappa_m$. We set $\sigma_1 = \kappa_1, \sigma_2 = \kappa_1 + \kappa_2, \dots$, and $\sigma_m = \kappa_1 + \kappa_2 + \dots + \kappa_m = n$.

Theorem 4.3: If the system (1) is T -related to the system (4) on U , then

- $\partial T_j / \partial u_k = 0$, $j = 1, 2, \dots, n$ and $k = 1, 2, \dots, m$,
- the $m \times m$ matrix $\{\partial T_j / \partial u_k\}$, $j = n+1, n+2, \dots, n+m$ and $k = 1, 2, \dots, m$ is nonsingular on U ,
- the following partial differential equations hold on U .

$$\begin{aligned}
\langle dT_{\ell}, g_i \rangle &= 0, \ell = 1, 2, \dots, \sigma_1 - 1, \sigma_1 + 1, \dots, \\
&\quad \sigma_2 - 1, \sigma_2 + 1, \dots, \sigma_{m-1} - 1, \sigma_{m-1} + 1, \\
&\quad \dots, n - 1 \text{ and } i = 1, 2, \dots, m, \\
\langle dT_{\ell}, f \rangle &= T_{\ell+1}, \ell = 1, 2, \dots, \sigma_1 - 1, \sigma_1 + 1, \dots, \\
&\quad \sigma_2 - 1, \sigma_2 + 1, \dots, \sigma_{m-1} - 1, \sigma_{m-1} + 1, \\
&\quad \dots, n - 1. \\
\langle dT_{\sigma_1}, f + \sum_{i=1}^m u_i g_i \rangle &= T_{n+1}, \\
\langle dT_{\sigma_2}, f + \sum_{i=1}^m u_i g_i \rangle &= T_{n+2}, \\
&\quad \vdots \\
\langle dT_{\sigma_m}, f + \sum_{i=1}^m u_i g_i \rangle &= T_{n+m}.
\end{aligned} \tag{5}$$

Proof. With x_0, u, y_0, v, ϕ and ψ as in Definition 3.2, let $y_j(t) = T_j(\phi(t; x_0, u), u(t))$ for $j = 1, 2, \dots, n$ and $v_j = T_{n+j}(\phi(t; x_0, u), u(t))$ for $j = 1, 2, \dots, m$. By hypothesis y_j is a state vector of system (4) with respect to $v(t)$ and

$$\frac{\partial y_j}{\partial t} = \sum_{k=1}^n \frac{\partial T_j}{\partial x_k} \frac{\partial x_k}{\partial t} + \sum_{k=1}^m \frac{\partial T_j}{\partial u_k} \frac{\partial u_k}{\partial t}, \quad j = 1, 2, \dots, n.$$

Since the rows of $\hat{A}y + \hat{B}v$ in (4) are independent of $\partial u_k / \partial t$, $k = 1, 2, \dots, m$, then $\partial T_j / \partial u_k = 0$, $j = 1, 2, \dots, n$ and $k = 1, 2, \dots, m$. Because the transformation T is a diffeomorphism by definition the matrix $\{\partial T_j / \partial u_k\}$, $j = n+1, n+2, \dots, n+m$ and $k = 1, 2, \dots, m$ is nonsingular on U .

From the canonical form (4) we have

$$\begin{aligned}
\hat{T}_{\ell} &= T_{\ell+1}, \ell = 1, 2, \dots, \sigma_1 - 1, \sigma_1 + 1, \dots, \sigma_2 - 1, \sigma_2 + 1, \dots, \\
&\quad \sigma_{m-1} - 1, \sigma_{m-1} + 1, \dots, n - 1,
\end{aligned}$$

which is equivalent to

$$\langle dT_{\ell}, f + \sum_{i=1}^m u_i g_i \rangle = T_{\ell+1}$$

for the same set of i . Since T_1, T_2, \dots, T_n are independent of u_1, u_2, \dots, u_m we find

$$\begin{aligned} \langle dT_i, g_i \rangle &= 0, i = 1, 2, \dots, \sigma_1 - 1, \sigma_1 + 1, \dots, \sigma_2 - 1, \sigma_2 \\ &\quad + 1, \dots, \sigma_{m-1} - 1, \sigma_{m-1} + 1, \dots, n - 1, \text{ and} \\ i &= 1, 2, \dots, m, \end{aligned}$$

and

$$\begin{aligned} \langle dT_{i+1}, f \rangle &= T_{i+1}, i = 1, 2, \dots, \sigma_1 - 1, \sigma_1 + 1, \dots, \sigma_2 - 1, \sigma_2 \\ &\quad + 1, \dots, \sigma_{m-1} - 1, \sigma_{m-1} + 1, \dots, n - 1. \end{aligned}$$

The canonical form (4) also yields

$$\dot{T}_{\sigma_1} = \langle dT_{\sigma_1}, f + \sum_{i=1}^m u_i g_i \rangle = T_{n+1},$$

$$\dot{T}_{\sigma_2} = \langle dT_{\sigma_2}, f + \sum_{i=1}^m u_i g_i \rangle = T_{n+2},$$

\vdots

$$\dot{T}_n = \langle dT_n, f + \sum_{i=1}^m u_i g_i \rangle = T_{n+m},$$

and the partial differential equations hold for T . \square

Note that

$$\begin{aligned} \langle dT_i, f \rangle &= T_{i+1} = L_f(T_i), i = 1, 2, \dots, \sigma_1 - 1, \sigma_1 + 1, \dots, \sigma_2 - 1, \\ &\quad \sigma_2 + 1, \dots, \sigma_{m-1} - 1, \sigma_{m-1} + 1, \dots, n - 1. \end{aligned}$$

From this and the Leibnitz formula (2) we deduce $\langle dT_2, g_1 \rangle = 0$ implies $\langle dT_1, [f, g_1] \rangle = 0$ for $i = 1, 2, \dots, m$, and $\langle dT_3, g_1 \rangle = 0$ implies $\langle dT_1, (ad^2 f, g_1) \rangle = 0$ for $i = 1, 2, \dots, m$, etc. Hence, the partial differential equations (5) become

$$\left. \begin{aligned} \langle dT_1, (ad^j f, g_i) \rangle &= 0, j = 0, 1, \dots, \kappa_1 - 2 \text{ and} \\ i &= 1, 2, \dots, m, \\ \langle dT_{\sigma_1+1}, (ad^j f, g_i) \rangle &= 0, j = 0, 1, \dots, \kappa_2 - 2 \text{ and} \\ i &= 1, 2, \dots, m, \\ &\vdots \end{aligned} \right\} (6)$$

$$\langle dT_{\sigma_{m-1}+1}, (ad^j f, g_i) \rangle = 0, \quad j = 0, 1, \dots, \kappa_m - 2 \quad \text{and} \\ i = 1, 2, \dots, m,$$

$$\langle dT_{\sigma_1}, f + \sum_{i=1}^m u_i g_i \rangle = T_{n+1},$$

$$\langle dT_{\sigma_2}, f + \sum_{i=1}^m u_i g_i \rangle = T_{n+2},$$

⋮

$$\langle dT_n, f + \sum_{i=1}^m u_i g_i \rangle = T_{n+m}.$$

(6)
Concl'd.

Since

$$T_{\sigma_1} = L_f(T_{\sigma_1-1})$$

$$T_{\sigma_2} = L_f(T_{\sigma_2-1})$$

⋮

$$T_n = L_f(T_{n-1})$$

the last m equations in (6) by the Liebnitz formula become

$$\langle dT_{\sigma_1}, f \rangle \pm \sum_{i=1}^m u_i \langle dT_1, (ad^{\kappa_1-1} f, g_i) \rangle = T_{n+1},$$

$$\langle dT_{\sigma_2}, f \rangle \pm \sum_{i=1}^m u_i \langle dT_{\sigma_1+1}, (ad^{\kappa_2-1} f, g_i) \rangle = T_{n+2},$$

⋮

$$\langle dT_n, f \rangle \pm \sum_{i=1}^m u_i \langle dT_{\sigma_{m-1}+1}, (ad^{\kappa_m-1} f, g_i) \rangle = T_{n+m},$$

where $+$ is for κ_i odd and $-$ is for κ_i even, $i = 1, 2, \dots, m$.

In order to solve for u_1, u_2, \dots, u_m in these last equations the matrix

$$\begin{bmatrix} \langle dT_1, (ad^{\kappa_1-1} f, g_1) \rangle & \dots & \langle dT_1, (ad^{\kappa_1-1} f, g_m) \rangle \\ \langle dT_{\sigma_1+1}, (ad^{\kappa_2-1} f, g_1) \rangle & \dots & \langle dT_{\sigma_1+1}, (ad^{\kappa_2-1} f, g_m) \rangle \\ \vdots & & \vdots \\ \langle dT_{\sigma_{m-1}+1}, (ad^{\kappa_m-1} f, g_1) \rangle & \dots & \langle dT_{\sigma_{m-1}+1}, (ad^{\kappa_m-1} f, g_m) \rangle \end{bmatrix}$$

must be nonsingular.

Hence, the partial differential equations that we solve to construct a transformation from system (1) to system (4) are

$$\begin{aligned}
 & \langle dT_1, (\text{ad}^j f, g_i) \rangle = 0, \quad j = 0, 1, \dots, \kappa_1 - 2 \quad \text{and} \\
 & \quad i = 1, 2, \dots, m, \\
 & \langle dT_{\sigma_1+1}, (\text{ad}^j f, g_i) \rangle = 0, \quad j = 0, 1, \dots, \kappa_2 - 2 \quad \text{and} \\
 & \quad i = 1, 2, \dots, m, \\
 & \quad \vdots \\
 & \langle dT_{\sigma_{m-1}+1}, (\text{ad}^j f, g_i) \rangle = 0, \quad j = 0, 1, \dots, \kappa_m - 2 \quad \text{and} \\
 & \quad i = 1, 2, \dots, m, \\
 & \langle dT_{\sigma_1}, f \rangle = \sum_{i=1}^m u_i \langle dT_1, (\text{ad}^{\kappa_1-1} f, g_i) \rangle = T_{n+1}, \\
 & \langle dT_{\sigma_2}, f \rangle = \sum_{i=1}^m u_i \langle dT_{\sigma_1+1}, (\text{ad}^{\kappa_2-1} f, g_i) \rangle = T_{n+2}, \\
 & \quad \vdots \\
 & \langle dT_n, f \rangle = \sum_{i=1}^m u_i \langle dT_{\sigma_{m-1}+1}, (\text{ad}^{\kappa_m-1} f, g_i) \rangle = T_{n+m}.
 \end{aligned} \tag{7}$$

with the determinant of

$$\begin{bmatrix}
 \langle dT_1, (\text{ad}^{\kappa_1-1} f, g_1) \rangle & \dots & \langle dT_1, (\text{ad}^{\kappa_1-1} f, g_m) \rangle \\
 \langle dT_{\sigma_1+1}, (\text{ad}^{\kappa_2-1} f, g_1) \rangle & \dots & \langle dT_{\sigma_1+1}, (\text{ad}^{\kappa_2-1} f, g_m) \rangle \\
 \vdots & & \vdots \\
 \langle dT_{\sigma_{m-1}+1}, (\text{ad}^{\kappa_m-1} f, g_1) \rangle & \dots & \langle dT_{\sigma_{m-1}+1}, (\text{ad}^{\kappa_m-1} f, g_m) \rangle
 \end{bmatrix} \tag{7'}$$

being nonzero.

The existence of solutions to (7) depends on the Frobenius Theorem mentioned earlier.

5. Existence of Transformations.

Recall we defined the following sets.

$$C = \{g_1, [f, g_1], \dots, (ad^{K_1-1} f, g_1), g_2, [f, g_2], \dots, (ad^{K_2-1} f, g_2), \dots, g_m, [f, g_m], \dots, (ad^{K_m-1} f, g_m)\}.$$

$$C_j = \{g_1, [f, g_1], \dots, (ad^{K_j-2} f, g_1), g_2, [f, g_2], \dots, (ad^{K_j-2} f, g_2), \dots, g_m, [f, g_m], \dots, (ad^{K_j-2} f, g_m)\} \text{ for } j = 1, 2, \dots, m,$$

We want to show that our system (1) is T-equivalent to the system (4) near the origin if and only if

i) the set C spans an n dimensional space (i.e., the vectors in C are linearly independent),

ii) the sets C_j are involutive for $j = 1, 2, \dots, m$ and

iii) the span of C_j equals the span of $C_j \cap C$ for $j = 1, 2, \dots, m$,

all on some neighborhood of the origin in \mathbb{R}^n . First, we show that condition i) is necessary and sufficient for $dT_1, dT_2, \dots, dT_{n+m}$ to be linearly independent and conditions ii) and iii) are necessary in order that a T-transformation exists. A constructive proof will be given for the sufficiency of conditions ii) and iii).

We compare the linear independence of the set of vector fields in C with the linear independence of the gradients dT_1, dT_2, \dots, dT_n assuming that T_1, T_2, \dots, T_{n+m} solve equations (7). Let c_1, c_2, \dots, c_n be constants and form the vector field

$$\begin{aligned} \alpha = & c_1 g_1 + c_2 [f, g_1] + \dots + c_{\sigma_1} (ad^{K_1-1} f, g_1) + c_{\sigma_1+1} g_2 \\ & + c_{\sigma_1+2} [f, g_2] + \dots + c_{\sigma_2} (ad^{K_2-1} f, g_2) + \dots + c_{\sigma_{m-1}+1} g_m \\ & + c_{\sigma_{m-1}+2} [f, g_m] + \dots + c_n (ad^{K_m-1} f, g_m) = 0. \end{aligned}$$

Taking the operation $\langle \cdot, \alpha \rangle$ with each dT_1, dT_2, \dots, dT_n and applying equations (7) and the Leibnitz formula (2) until only $dT_1, dT_{\sigma_1+1}, \dots, dT_{\sigma_{m-1}+1}$ are involved, we find that the vector

fields in C are linearly independent if and only if the determinant of the matrix

$$\begin{bmatrix} \langle dT_1, (ad^{K_1-1} f, g_1) \rangle & \dots & \langle dT_1, (ad^{K_1-1} f, g_m) \rangle \\ \langle dT_{\sigma_1+1}, (ad^{K_2-1} f, g_1) \rangle & \dots & \langle dT_{\sigma_1+1}, (ad^{K_2-1} f, g_m) \rangle \\ \vdots & & \vdots \\ \langle dT_{\sigma_{m-1}+1}, (ad^{K_m-1} f, g_1) \rangle & \dots & \langle dT_{\sigma_{m-1}+1}, (ad^{K_m-1} f, g_m) \rangle \end{bmatrix}$$

is nonzero.

Similarly, with constants b_1, b_2, \dots, b_n we define the one form

$$\beta = b_1 dT_1 + b_2 dT_2 + \dots + b_n dT_n = 0.$$

Taking the operation $\langle \beta, \cdot \rangle$ with each vector field in C , and applying equations (7) and the Leibnitz formula (2) until only $dT_1, dT_{\sigma_1+1}, \dots, dT_{\sigma_{m-1}+1}$ are involved, we find that dT_1, dT_2, \dots, dT_n are linearly independent if and only if the determinant of the matrix

$$\begin{bmatrix} \langle dT_1, (ad^{K_1-1} f, g_1) \rangle & \langle dT_{\sigma_1+1}, (ad^{K_2-1} f, g_1) \rangle & \dots & \langle dT_{\sigma_{m-1}+1}, (ad^{K_m-1} f, g_1) \rangle \\ \vdots & \vdots & & \vdots \\ \langle dT_1, (ad^{K_1-1} f, g_m) \rangle & \langle dT_{\sigma_1+1}, (ad^{K_2-1} f, g_m) \rangle & \dots & \langle dT_{\sigma_{m-1}+1}, (ad^{K_m-1} f, g_m) \rangle \end{bmatrix}$$

is nonvanishing.

Thus, for T_1, T_2, \dots, T_{n+m} satisfying (7), $dT_1, dT_2, \dots, dT_{n+m}$ are linearly independent if and only if the vector fields in C are linearly independent if and only if the determinant of the matrix in (7') is nonzero. We have used the fact that in $T = (T_1, T_2, \dots, T_n, T_{n+1}, \dots, T_{n+m})$ only $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ are functions of u_1, u_2, \dots, u_m . Also we have shown that to find a T -relation between systems (1) and (4) we need to solve equations (7) with the matrix in (7') being nonsingular.

Next, we examine conditions ii) and iii) under the assumption that dT_1, dT_2, \dots, dT_n are linearly independent (i.e., under the equivalent assumption i)). The equations $\langle dT_1, (ad^j f, g_i) \rangle = 0$, $j = 0, 1, \dots, \kappa_1 - 2$ and $i = 1, 2, \dots, m$ imply that the vector fields in C_1 span at most an $n - 1$ dimensional space at each point. If $\kappa_1 > \kappa_2$ we have that the span of C_1 equals the span of $C_1 \cap C$, since the vector fields in C are linearly independent. Also we have that the set C_1 (thus, the set $C_1 \cap C$) must be involutive with dT_1 being a nonvanishing normal to the $n - 1$ dimensional integral manifold of $C_1 \cap C$ given by the Frobenius Theorem. If $\kappa_1 = \kappa_2 = \dots = \kappa_s$ (i.e., we say κ_1 appears in C s times), for some integer $s \geq 2$, then the vector fields in C_1 span at most an $n - s$ dimensional space at each point from equations (7) and the linear independence of $dT_1, dT_{\sigma_1+1}, \dots, dT_{\sigma_{s-1}+1}$. Again, because of the linear independence of vectors in C , the span of C_1 is equal to the span of $C_1 \cap C$. The set C_1 (and $C_1 \cap C$) is involutive with $dT_1, dT_{\sigma_1+1}, \dots, dT_{\sigma_{s-1}+1}$ being normal to the $n - s$ dimensional manifold of $C_1 \cap C$ given by Frobenius.

For our present purpose the interesting equations from (7) are

$$\begin{aligned}
 \langle dT_1, (ad^j f, g_i) \rangle &= 0, \quad j = 0, 1, \dots, \kappa_1 - 2 \quad \text{and} \\
 &\quad i = 1, 2, \dots, m, \\
 \langle dT_{\sigma_1+1}, (ad^j f, g_i) \rangle &= 0, \quad j = 0, 1, \dots, \kappa_2 - 2 \quad \text{and} \\
 &\quad i = 1, 2, \dots, m, \\
 &\vdots \\
 \langle dT_{\sigma_{m-1}+1}, (ad^j f, g_i) \rangle &= 0, \quad j = 0, 1, \dots, \kappa_m - 2 \quad \text{and} \\
 &\quad i = 1, 2, \dots, m.
 \end{aligned} \tag{8}$$

From the Leibnitz rule (2) we have

$$\langle dT_\ell, (ad^j f, g_i) \rangle = -\langle dT_{\ell+1}, (ad^{j-1} f, g_i) \rangle,$$

and this can be used repeatedly. Thus, given a j , $1 \leq j \leq m$, since dT_1, dT_2, \dots, dT_n are linearly independent, the number of linearly independent vector fields in C_j is $n - (p/m)$, where p is the number of equations in (8) for which the k in $(ad^k f, g_i)$ is greater than or equal to $\kappa_j - 2$. This is exactly the number of linearly independent vector fields in $C_j \cap C$. From the set T_1, T_2, \dots, T_n

there are $n - [(n - (p/m))] = p/m$ functions with linearly independent gradients that are normal to the integral manifold of $C_j \cap C$ (or equivalently of C_j).

Hence, the necessity of conditions ii) and iii) are proved. We now state our main result and complete its proof. We must keep in mind that a reordering of the vector fields g_1, g_2, \dots, g_m has been done, if necessary.

Theorem 5.1: The system (1) is T-equivalent to the system (4), where the state variables x_1, x_2, \dots, x_n lie in a sufficiently small open neighborhood V of the origin in \mathbb{R}^n , if and only if conditions i), ii), and iii) hold on V .

Proof. The necessary part of the theorem has been proven. Thus, we wish to construct a T-transformation given the three conditions and remember that our transformation will be nonsingular by i). Since the span of C_j equals the span of $C_j \cap C$ for $j = 1, 2, \dots, m$, we have that each $C_j \cap C$ is involutive and work with only those vector fields in C .

Before we build T we assume that the following conditions hold. Suppose all entries when evaluated at the origin in the matrix (7') above the diagonal are zero. After the proof is completed we will show that no generality is lost in making this assumption. Hence, the T_1, T_2, \dots, T_{n+m} that we define have linearly independent gradients and the matrix in (7') is nonsingular if and only if the diagonal elements in this matrix do not vanish.

We now construct a solution $T = (T_1, T_2, \dots, T_{n+m})$ of equations (7). Since $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ are found by the last m equations in (7), we need only find solutions $T_1, T_{\sigma_1+1}, \dots, T_{\sigma_{m-1}+1}$ to the partial differential equations (8). Recall that first-order linear partial differential equations are solved by reducing to systems of ordinary differential equations. We now give the ordinary differential equations to be examined and then choose our solutions $T_1, T_{\sigma_1+1}, \dots, T_{\sigma_{m-1}+1}$. It should be evident that these m functions will not be unique.

Let s_1 be the number of times that $\kappa_1 - 1$ appears in C (e.g., if $\kappa_1 = \kappa_2 > \kappa_3$, then $s_1 = 2$); s_2 be the number of times $\kappa_1 - 2$ appears in C , \dots , s_{κ_1} be the number of linearly independent vectors in $\{g_1, g_2, \dots, g_m\}$, which by assumption is m .

We introduce real parameters t_1, t_2, \dots, t_n in the following technique. We first solve

$$\frac{dx(t_1)}{dt_1} = (\text{ad}^{K_1-1} f, g_1)$$

with initial conditions $x(0) = 0$ to find the unique integral curve of $(\text{ad}^{K_1-1} f, g_1)$ through the origin in (x_1, x_2, \dots, x_n) space. If $s_1 \geq 2$, we next solve the system

$$\frac{dx(t_1, t_2)}{dt_2} = (\text{ad}^{K_1-1} f, g_2)$$

with initial conditions $x(t_1, 0) = x(t_1)$ to find the solution with $x(0, 0) = 0$. We continue in this manner $s_1 - 2$ more times finally finding the solution of

$$\frac{dx(t_1, t_2, \dots, t_{s_1})}{dt_{s_1}} = (\text{ad}^{K_1-1} f, g_{s_1})$$

satisfying $x(t_1, t_2, \dots, t_{s_1-1}, 0) = x(t_1, t_2, \dots, t_{s_1-1})$.

We now approach the partial differential equations (8). Finding a solution z to the equation $\langle dz, (\text{ad}^{K_1-2} f, g_1) \rangle = 0$ is the same as solving

$$\frac{dx}{dt_{s_1+1}} = (\text{ad}^{K_1-2} f, g_1), \quad \frac{\partial z}{\partial t_{s_1+1}} = 0$$

with initial conditions $x(t_1, t_2, \dots, t_{s_1}, 0) = x(t_1, t_2, \dots, t_{s_1})$.

We let z denote the unknown function for each equation we consider.

If $s_2 > 1$ for $\langle dz, (\text{ad}^{K_1-2} f, g_2) \rangle = 0$ we examine

$$\frac{dx}{dt_{s_1+2}} = (\text{ad}^{K_1-2} f, g_2), \quad \frac{\partial z}{\partial t_{s_1+2}} = 0$$

satisfying $x(t_1, t_2, \dots, t_{s_1+1}, 0) = x(t_1, t_2, \dots, t_{s_1+1})$. Doing this s_2 times takes us through the partial differentiation equation $\langle dz, (\text{ad}^{K_1-2} f, g_{s_2}) \rangle = 0$ and parameter $t_{s_1+s_2}$.

Next we consider $\langle dz, (\text{ad}^{K_1-3} f, g_1) \rangle = 0$ and take the system

$$\frac{dx}{dt_{s_1+s_2+1}} = (\text{ad}^{K_1-3} f, g_1), \quad \frac{\partial z}{\partial t_{s_1+s_2+1}} = 0$$

with initial conditions $x(t_1, t_2, \dots, t_{s_1+s_2}, 0)$
 $= x(t_1, t_2, \dots, t_{s_1+s_2})$. We repeat this process $s_3 - 1$ more times
 if $s_3 > 1$ and end with the system

$$\frac{dx}{dt_{s_1+s_2+s_3}} = (\text{ad}^{k_1-3} f, g_{s_3}), \quad \frac{\partial z}{\partial t_{s_1+s_2+s_3}} = 0$$

and parameter $t_{s_1+s_2+s_3}$ associated with $\langle dz, (\text{ad}^{k_1-3} f, g_{s_3}) \rangle = 0$.

Then we examine $\langle dz, (\text{ad}^{k_1-4} f, g_1) \rangle = 0$ and continue this process
 in order, introducing all parameters t_1, t_2, \dots, t_n and ending with
 the equations (recall $s_1 + s_2 + \dots + s_{k_1} = n$)

$$\frac{dx}{dt_n} = g_m, \quad \frac{\partial z}{\partial t_n} = 0$$

satisfying $x(t_1, t_2, \dots, t_{n-1}, 0) = x(t_1, t_2, \dots, t_{n-1})$. In this
 way we treat $\langle dz, g_m \rangle = 0$.

We have constructed a mapping from \mathbb{R}^n to \mathbb{R}^n given by

$$(t_1, t_2, \dots, t_n) \mapsto (x_1(t_1, t_2, \dots, t_n), x_2(t_1, t_2, \dots, t_n), \dots, x_n(t_1, t_2, \dots, t_n))$$

and taking the origin to the origin. The Jacobian matrix of this
 mapping

$$\begin{bmatrix} \frac{\partial x_1}{\partial t_1} & \frac{\partial x_1}{\partial t_2} & \dots & \frac{\partial x_1}{\partial t_n} \\ \frac{\partial x_2}{\partial t_1} & \frac{\partial x_2}{\partial t_2} & \dots & \frac{\partial x_2}{\partial t_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial t_1} & \frac{\partial x_n}{\partial t_2} & \dots & \frac{\partial x_n}{\partial t_n} \end{bmatrix}$$

is called the noncharacteristic matrix. When evaluated at the origin,
 its columns are $(\text{ad}^{k_1-1} f, g_1), (\text{ad}^{k_2-1} f, g_2), \dots, [f, g_m], g_1, \dots, g_m$,
 the vector fields in C all evaluated at $(0, 0, \dots, 0)$. Thus, the
 matrix is nonsingular and the above mapping $(x_1(t_1, t_2, \dots, t_n),$
 $x_2(t_1, t_2, \dots, t_n), \dots, x_n(t_1, t_2, \dots, t_n))$ is a diffeomorphism
 on an open neighborhood W of the origin in \mathbb{R}^n . We let V be a
 sufficiently small open set about $(0, 0, \dots, 0)$ in the image of W

under the mapping. Moreover, we can solve for t_1, t_2, \dots, t_n as functions of x_1, x_2, \dots, x_n by the inverse function theorem.

We now define the functions $T_1, T_{\sigma_1+1}, \dots, T_{\sigma_{m-1}+1}$ to solve equations (8). By design, each maps the origin to the origin and is \mathcal{C}^∞ on V . Let $T_1 = t_1$ and note that $\partial T_1 / \partial t_i = 0$, $i = 2, 3, \dots, n$. We must show that $\langle dT_1, q \rangle = 0$ for every element q in $C_1 \cap C$ (this is the same as the equations involving T_1 in (8)). From the Frobenius Theorem we know there are integral manifolds of the vector fields in $C_1 \cap C$. By our construction, for fixed t_1, t_2, \dots, t_{s_1} as we let $t_{s_1+1}, t_{s_1+2}, \dots, t_n$ vary we obtain such an integral manifold. Thus, $T_1 = t_1$ is a constant on each such manifold and $\langle dT_1, q \rangle = 0$ for all q in $C_1 \cap C$. Note that t_1 is the parameter associated with the vector field $(\text{ad}^{K_1-1} f, g_1)$. Next, we define $T_{\sigma_1+1} = \bar{t}$, where \bar{t} is the parameter that was introduced when we solved $dx/d\bar{t} = (\text{ad}^{K_2-1} f, g_2)$. For example, if $s_1 = 2$, $T_{\sigma_1+1} = t_2 = \bar{t}$, or if $s_1 = 1$, $T_{\sigma_1+1} = t_3 = \bar{t}$. We must show that $\langle dT_{\sigma_1+1}, q \rangle = 0$ for all vector fields q in $C_2 \cap C$. Since $C_2 \cap C$ is involutive, there is a subdivision of the set of indices t_1, t_2, \dots, t_n into 2 subsets $\bar{S} = \{t_1, t_2, \dots, t_r\}$ and $\hat{S} = \{t_{r+1}, t_{r+2}, \dots, t_n\}$ for some r so that $\bar{t} \in \bar{S}$ and for fixed parameters in \bar{S} we get integral manifolds of $C_2 \cap C$ when $t_{r+1}, t_{r+2}, \dots, t_n$ vary. Thus, T_{σ_1+1} is constant on each such manifold and $\langle dT_{\sigma_1+1}, q \rangle = 0$ for all $q \in C_2 \cap C$.

We define T_{σ_2+1} to be the parameter associated with the vector field $(\text{ad}^{K_3-1} f, g_3)$, and continuing in this manner, we introduce $T_{\sigma_3+1}, T_{\sigma_4+1}, \dots, T_{\sigma_{m-1}+1}$ to solve (8). Also, we have $\langle dT_1, (\text{ad}^{K_1-1} f, g_1) \rangle \neq 0$, $\langle dT_{\sigma_1+1}, (\text{ad}^{K_2-1} f, g_2) \rangle \neq 0, \dots$, $\langle dT_{\sigma_{m-1}+1}, (\text{ad}^{K_m-1} f, g_m) \rangle \neq 0$. By a comment near the beginning of this proof, we know that our transformation is nonsingular because the diagonal elements in (7') are nonzero. \square

We point out how to redefine g_1, g_2, \dots, g_m , if necessary, so that each entry evaluated at the origin in (7') above the diagonal is zero. This technique does not alter the hypotheses of Theorem 4.1.

We take $g_1 = g_1$ and replace g_2 by $g_2 - e_{12}g_1$, where e_{12} is a constant so that the vector fields $(\text{ad}^{K_1-1} f, g_1)$ and $(\text{ad}^{K_1-1} f, g_1 - e_{12}g_2)$ have zero inner product at the origin. We call this $g_2 - e_{12}g_1$ our new g_2 . Next we replace g_3 by $g_3 - e_{13}g_1 - e_{23}g_2$, where e_{13} and e_{23} are chosen so that the inner

products of $(\text{ad}^{k_1-1}f, g_1)$ with $(\text{ad}^{k_1-1}f, g_3 - e_{13}g_1 - e_{23}g_2)$ and $(\text{ad}^{k_2-1}f, g_2)$ with $(\text{ad}^{k_2-1}f, g_3 - e_{13}g_1 - e_{23}g_2)$ vanish at $(0, 0, \dots, 0)$. This $g_3 - e_{13}g_1 - e_{23}g_2$ becomes our new g_3 . We continue in this way, the last step being to take $e_{1m}, e_{2m}, \dots, e_{m-1,m}$ to make the obvious vector fields orthogonal at the origin, and $g_m - e_{1m}g_1 - e_{2m}g_2 - \dots - e_{m-1,m}g_{m-1}$ is our new g_m .

Since $T_1 = t_1$, $dx/dt_1 = (\text{ad}^{k_1-1}f, g_1)$, we have $\langle dT_1, (\text{ad}^{k_1-1}f, g_2) \rangle = 0$, $\langle dT_1, (\text{ad}^{k_1-1}f, g_3) \rangle = 0, \dots$, $\langle dT_1, (\text{ad}^{k_1-1}f, g_m) \rangle = 0$, all evaluated at the origin. Because $T_{\sigma_1+1} = \tilde{t}$, $dx/d\tilde{t} = (\text{ad}^{k_2-1}f, g_2)$, we find that $\langle dT_{\sigma_1+1}, (\text{ad}^{k_2-1}f, g_3) \rangle = 0$, $\langle dT_{\sigma_1+1}, (\text{ad}^{k_2-1}f, g_4) \rangle = 0, \dots, \langle dT_{\sigma_1+1}, (\text{ad}^{k_2-1}f, g_m) \rangle = 0$ when $(x_1, x_2, \dots, x_n) = (0, 0, \dots, 0)$. Repeating this argument $m-3$ more times we show the entries in (7') above the diagonal vanish at the origin.

An illustrative example is in order.

Example 5.2: Consider the nonlinear system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \begin{bmatrix} \sin x_2 \\ \sin x_3 \\ x_4^3 \\ x_5 + x_4^3 - x_1^{10} \\ 0 \end{bmatrix} + u_1 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$$= f + u_1 g_1 + u_2 g_2$$

on \mathbb{R}^5 .

Now we compute

$$[f, g_1] = \begin{bmatrix} 0 \\ -\cos x_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad (\text{ad}^2 f, g_1) = \begin{bmatrix} \cos x_2 \cos x_3 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad [f, g_2] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Hence, $C = \{g_1, [f, g_1], (\text{ad}^2 f, g_1), g_2, [f, g_2]\}$ spans a 5-dimensional space on the set

$$V = \{(x_1, x_2, x_3, x_4, x_5) : -\frac{\pi}{2} < x_2, x_3 < \frac{\pi}{2}\}.$$

The appropriate Kronecker indices in this case are $\kappa_1 = 3$ and $\kappa_2 = 2$.

We have $C_1 \cap C = \{g_1, [f, g_1], g_2, [f, g_2]\}$ is involutive since

$$[g_1, [f, g_1]] = \begin{bmatrix} 0 \\ \sin x_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and all other Lie brackets vanish, and $C_2 \cap C = \{g_1, g_2\}$ is trivially involutive. Hence, there is a transformation which maps our nonlinear system to the appropriate Brunovsky canonical form.

We exhibit such a transformation for $(x_1, x_2, x_3, x_4, x_5)$ in V by the construction in the proof of Theorem 5.1. The solution of $dx/dt_1 = (\text{ad}^2 f, g_1)$ with $x(0) = 0$ is $x_1(t_1) = t_1, x_2(t_1) = 0, x_3(t_1) = 0, x_4(t_1) = 0$, and $x_5(t_1) = 0$. Similarly, for $dx/dt_2 = [f, g_1]$ with $x(t_1, 0) = x(t_1)$ we find $x_1(t_1, t_2) = t_1, x_2(t_1, t_2) = -t_2, x_3(t_1, t_2) = 0, x_4(t_1, t_2) = 0$, and $x_5(t_1, t_2) = 0$. For $dx/dt_3 = [f, g_2]$, $dx/dt_4 = g_1$, and $dx/dt_5 = g_2$ in that order and with the appropriate initial conditions we have $x_1(t_1, t_2, t_3, t_4, t_5) = t_1, x_2(t_1, t_2, t_3, t_4, t_5) = -t_2, x_3(t_1, t_2, t_3, t_4, t_5) = t_4, x_4(t_1, t_2, t_3, t_4, t_5) = t_3$, and $x_5(t_1, t_2, t_3, t_4, t_5) = t_5$. Certainly, the noncharacteristic matrix is nonsingular and solving for t_1, t_2, t_3, t_4, t_5 as functions of x_1, x_2, x_3, x_4, x_5 we find $t_1 = x_1, t_2 = -x_2, t_3 = x_4, t_4 = x_3$, and $t_5 = x_5$.

Then one such transformation $T = (T_1, T_2, \dots, T_n)$ is

$$T_1 = x_1$$

$$T_2 = \sin x_2$$

$$T_3 = (\cos x_2) \sin x_3$$

$$T_4 = x_4$$

$$T_5 = x_5 + x_4^3 - x_1^{10}$$

$$T_6 = (-\sin x_2) \sin^2 x_3 + (\cos x_2 \cos x_3)(x_1 + x_4^2)$$

$$T_7 = x_2 + 3x_4^2(x_5 + x_4^3) - 10x_1^9 (\sin x_2).$$

Given the system (1) we choose the Kronecker indices in the following way. First we form the matrix

$$\begin{bmatrix} g_1 & g_2 & \dots & g_m \\ [f, g_1] & [f, g_2] & \dots & [f, g_m] \\ (\text{ad}^2 f, g_1) & (\text{ad}^2 f, g_2) & \dots & (\text{ad}^2 f, g_m) \\ \vdots & \vdots & & \vdots \\ (\text{ad}^{n-1} f, g_1) & (\text{ad}^{n-1} f, g_2) & \dots & (\text{ad}^{n-1} f, g_m) \end{bmatrix}$$

We assume that all numbers we define are constant in some neighborhood of the origin in \mathbb{R}^n . Set α_0 = number of linearly independent vector fields in the first row; α_1 = number of linearly independent vector fields in the first two rows, . . .; and α_{n-1} = number of linearly independent vector fields in the matrix. Take $r_0 = \alpha_0$, $r_1 = \alpha_1 - \alpha_0$, . . .; $r_{n-1} = \alpha_{n-1} - \alpha_{n-2}$ and define κ_j to be the number of r_j with $r_j \geq 1$.

If necessary, we renumber g_1, g_2, \dots, g_m and see if the hypotheses of Theorem 5.1 are satisfied to determine if our system is transformable.

In [3] the authors blended the local results for the single input case with versions of the global inverse function theorem to yield global transformation results. Of course analogous global results can be derived for the multiple input case with the local theory found in this paper.

Another problem on which the authors have made progress is the existence of transformations to time-invariant linear systems for time-varying nonlinear systems (see [4]). We can also show that if a nonlinear system can be mapped to a linear system, then it must be "reducible" to the form (1).

A conversation with Roger Brockett, Harvard University, on the multi-input transformation problem proved valuable to the authors. We wish to thank Eduardo Sontag, Rutgers University, for sending us a copy of the paper [5].

Appendix

APPLICATION TO A HELICOPTER

The helicopter is represented by a rigid body moving in 3-dimensional space in response to gravity, aerodynamics, and propulsion. The state,

$$x = (r, v, C, \omega)^T \in X \subset \mathbb{R}^3 \times \mathbb{R}^3 \times SO(3) \times \mathbb{R}^3 \quad (A-1)$$

where r and v are the inertial-coordinates of body center of mass position and velocity, respectively; C is the direction cosine matrix of the body-fixed axes relative to the runway-fixed axes (taken to be inertial). The attitude C moves in the Lie group $SO(3)$. The body coordinates of angular velocity are represented by ω .

The controls,

$$u = (u^M, u^P)^T \in U \subset \mathbb{R}^3 \times \mathbb{R} \quad (A-2)$$

where u^M is the 3-axis moment control, that is, roll cyclic and pitch cyclic, which tilt the main rotor thrust, and the tail rotor collective, which controls the yaw moment; and u^P is the main rotor collective, which controls the main motor thrust.

The effectively 12-dimensional state equation consists of the translational and rotational kinematic and dynamic equations:

$$\begin{aligned} \dot{r} &= v \\ \dot{v} &= f^F(x, u) \\ \dot{C} &= S(\omega)C \\ \dot{\omega} &= f^M(x, u) \end{aligned} \quad (A-3)$$

where f^F and f^M are the total force and moment generation processes, and (x, u) are defined by (A-1) and (A-2).

The canonic model has Kronecker indices $\{4,4,2,2\}$ and the variables are identified as shown in Figure A-1, where \bullet represents a scalar integrator, $y^5 \in \mathbb{R}^4$ is the canonic control and E_3 is the direction cosine matrix representing the heading of the helicopter.

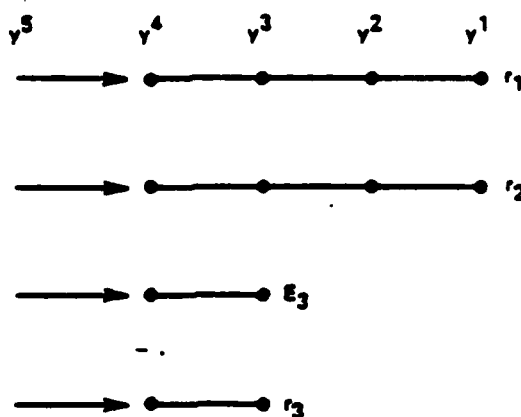


Figure A-1. Canonic Model of (A-3).

We do not go into details of the nonlinear system nor the construction of the linearizing transformation here. However, we mention that we have a method to build an approximate transformation and the accuracy of this technique is illustrated through the simulation results.

The experiment consists in automatically flying a trajectory which exercises most of the flight envelope of the UH-1H helicopter as shown in figures A-2 and A-3.

Unlike the coarse accelerations in figure A-3, the model accelerations (figure A-4) are smooth, as is the vertical velocity v_{30} . The second panel in figure A-4 shows the effects of neglected parasitic effects in the construction of the linearizing transformation. As can be seen the acceleration errors are quite small being less than 0.5 g. The regulator controls these effects by means of position errors. The resulting horizontal error is less than 6 ft; while the vertical error is below 1 ft.

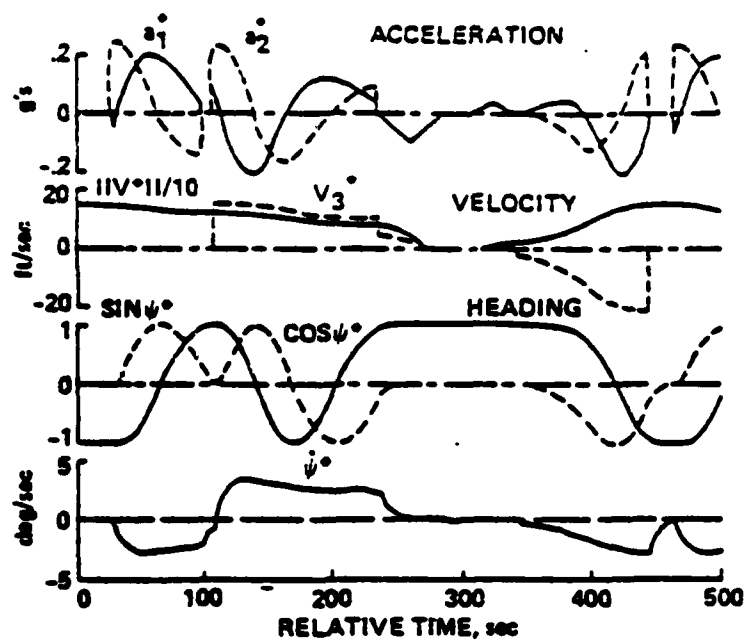


Figure A-3. Time Dependence of the Coarse Inputs.

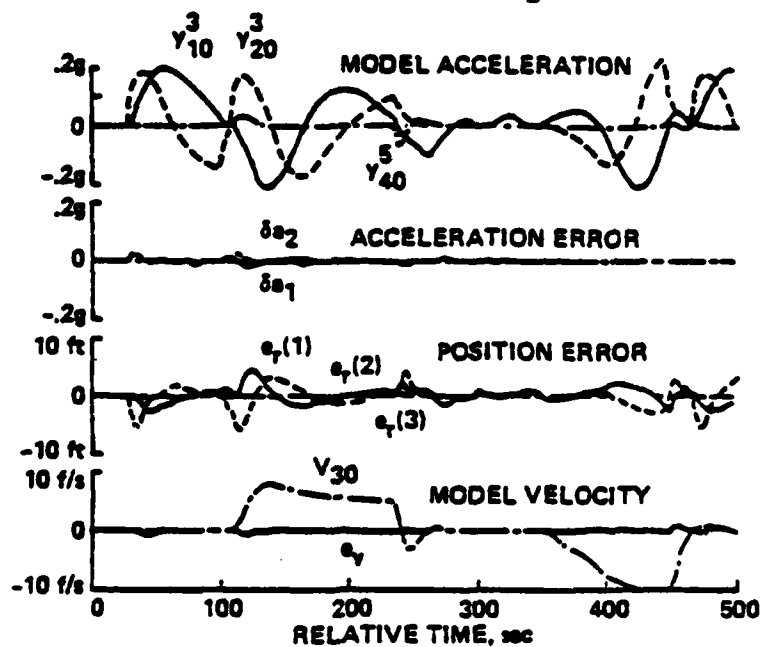


Figure A-4. System Response-Canonic Variables.

The natural controls $u = (u^M, u^P)$, shown in figure A-5 are well behaved.

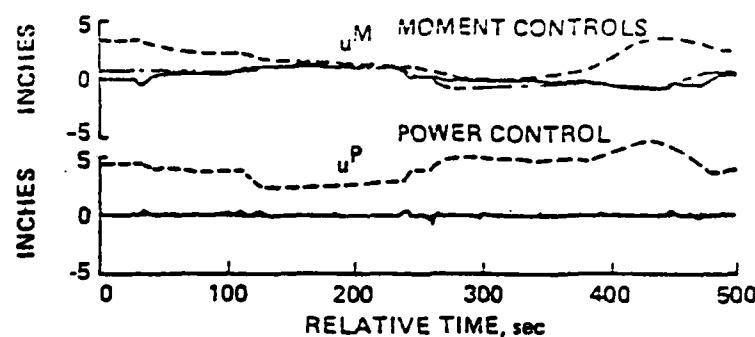
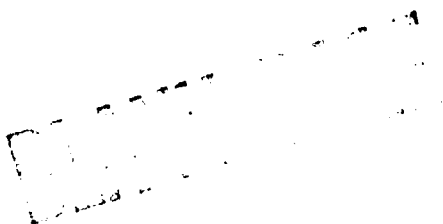


Figure A-5. Response Natural Controls.

In summary, the performance of the system is good. Future tests will exercise the system with more taxing trajectories.



REFERENCES

- [1] Brockett, R. W., 1978. Feedback invariants for nonlinear systems. IFAC Congress, Helsinki, 1978.
- [2] Brunovsky, P. A classification of linear controllable systems. Kibernetika (Prana) 6 (1970), pp. 173-186.
- [3] Hunt, L. R., Su, R. and Meyer, G. Global transformations of nonlinear systems. IEEE Trans. on Automatic Control, 27, No. 6 (1982), to appear.
- [4] Hunt, L. R. and Su, R. Linear equivalents of nonlinear time varying systems. International Symposium on Mathematical Theory of Networks and Systems, 1981, pp. 119-123.
- [5] Jakubczyk, B. and Respondek, W. On linearization of control systems. Bull. Acad. Polon. Sci., Ser. Sci. Math. Astronom. Phys., 28 (1980), pp. 517-522.
- [6] Kalman, R. E., 1972. Kronecker invariants and feedback, ordinary differential equations, L. Weiss, ed., Academic Press, pp. 459-471.
- [7] Krener, A. J. On the equivalence of control systems and the linearization of nonlinear systems. SIAM J. Control 11, 1973, pp. 670-676.
- [8] Meyer, G. and Cicolani, L., 1975. A formal structure for advanced automatic flight control systems. NASA TN D-7940.
- [9] Meyer, G. and Cicolani, L., 1980. Application of nonlinear system inverses to automatic flight control design - system concepts and flight evaluations. AGARDograph 251 on Theory and Applications of Optimal Control in Aerospace Systems, P. Kant, ed., reprinted by NATO.
- [10] Meyer, G. The design of exact nonlinear model followers, Proceedings of 1981 Joint Automatic Control Conference, FA3A.
- [11] Smith, G. A. and Meyer, G. Total aircraft flight control system balanced open- and closed-loop with dynamic trimmings, 3rd Avionics Conference, Dallas, 1979.
- [12] Smith, G. A. and Meyer, G. Applications of the concept of dynamic trim control to automatic landing of carrier aircraft, NASA TP-1512, 1980.
- [13] Su, R. On the linear equivalents of nonlinear systems, Systems and Control Letters 2, No. 1 (1982), to appear.
- [14] Wehrend, W. R., Jr. and Meyer, G. Flight tests of the total automatic flight control system (TAFCOS) concept on a DHC-6 Twin Otter aircraft. NASA TP-1513, 1980.
- [15] Wehrend, W. R., Jr. Pilot control through the TAFCOS automatic flight control system. NASA TM-81152, 1979.
- [16] Wonham, W. M. and Morse, A. S. Feedback invariants of linear multi-variable systems, Automatica 8, 1972, pp. 93-100.

Addresses of Authors

L. R. Hunt

Mail Stop 210-3

NASA Ames Research Center

Moffett Field, CA 94035

(415) 965-5453

on leave from

Department of Mathematics

Texas Tech University

Lubbock, Texas 79409

Renjeng Su

Mail Stop 210-3

NASA Ames Research Center

Moffett Field, CA 94035

(415) 965-5453

George Meyer

Mail Stop 210-3

NASA Ames Research Center

Moffett Field, CA 94035

(415) 965-5444

*Researcher supported by NASA Ames Research Center under the Inter-governmental Personnel Agreement Program and the Joint Services Electronics Program at Texas Tech University under Office of Naval Research Contract N00014-76-C-1136.

**National Research Council Research Associate at NASA Ames Research Center.

†Research Engineer at NASA Ames Research Center.

ROBUSTNESS IN- NONLINEAR CONTROL

Renjeng Su, George Meyer and L. R. Hunt

Abstract

A new design methodology for nonlinear plants using transformations of nonlinear systems to linear systems is presently being developed. It is the purpose of this paper to show that this design theory is robust. If the linear system is asymptotically stabilized by applying appropriate feedback (a well-known technique), then a control to stabilize the nonlinear plant is easily computed through that part of the inverse transformation involving controls. Most importantly, all nearby plants (in the proper topology) are also asymptotically stabilized using this control. Lyapunov functions for nonlinear systems can be found using this method. A short discussion on the application of this design technique to the automatic flight control of aircraft is presented.

ROBUSTNESS IN NONLINEAR CONTROL

Renjeng Su¹

George Meyer²

L. R. Hunt³

1. Introduction

Suppose we have a nonlinear system of the form $\dot{x}(t) = h(x, u)$ where h is a C^1 function of the state $x \in \mathbb{R}^n$ and of the (m -tuple) control vector u . If we choose a control vector u_0 and if $h(x, u_0)$ vanishes when $x = 0$, does the asymptotic stability of the zero solution imply the asymptotic stability of equilibrium points of nearby control systems (in an appropriate topology) with the control u_0 driving these systems? We work in some fixed set in \mathbb{R}^n which contains the origin.

In our version of this problem we have a nonlinear plant that we wish to control. The mathematical model of the plant is

$$\dot{x}(t) = f[x(t)] + \sum_{i=1}^m u_i(t) g_i[x(t)] \quad , \quad \dot{x} \equiv \frac{dx}{dt} \quad (1)$$

where f, g_1, \dots, g_m are vector fields which are C^∞ in an open set W in \mathbb{R}^n containing the origin and $f(0) = 0$. We assume

¹Research Associate of the National Research Council.

²Research Engineer at NASA Ames Research Center.

³Research supported by NASA Ames Research Center under the IPA Program and the Joint Services Electronics Program at Texas Tech University under ONR Contract N00014-76-C-1136.

that there is a nonsingular one-to-one transformation T mapping our system (1) to a controllable linear system in the appropriate Brunovsky canonical form [1]. Thus, our system (1) transforms by $T: W \times R^m \rightarrow T(W \times R^m) \subset R^n \times R^m$ with $T(0) = 0$ into

$$\dot{y}(t) = A_0 y + B_0 v \quad (2)$$

where y and v are the canonical state and control.

The design technique is to build a controller for the nonlinear system by designing one for the linear canonical system. For aircraft we use the canonical form to determine the control to fly a reference trajectory and to stabilize the system about this trajectory. Using the transformation (from nonlinear to linear) and its inverse we obtain an exact nonlinear model follower [2-4].

In this article we use linear feedback to asymptotically stabilize the linear system about 0, and this in turn asymptotically stabilizes the nonlinear system about the origin in its state space. We show that all systems close (in the topology we introduce) to the mathematical model are stabilized (asymptotically) about corresponding equilibrium points, and the stability holds for any trajectory starting in some fixed compact set in state space (of course, the usual linearization techniques are local in nature). In this way we prove that our design technique is robust. At the end of this paper we discuss applications of our approach to various aircraft and some flight test results which indicate the robustness of our technique in actual practice.

Lyapunov functions play a key role in our theory. If linear feedback is used to asymptotically stabilize the canonical system, then it is easy to construct Lyapunov functions. It is shown that the composition of these functions with the transformation yield Lyapunov functions for the nonlinear system

$$\dot{x}(t) = f[x(t)] + \sum_{i=1}^m u_i(t) g_i[x(t)]$$

(which we could also write as $\dot{x} = f + Gu$, G and u having obvious definitions) with the controls u_i , $i = 1, 2, \dots, m$, corresponding

to the linear feedback applied. If a nonlinear system is sufficiently close to the above system (in the C^1 topology on a particular compact set containing the origin), we achieve asymptotic stability for this nearby system.

We emphasize linear feedback on the linear system, but the technique can be applied for many asymptotically stabilizing controls for which we can find a Lyapunov function whose level sets (sets where this function is constant) shrink to the origin in the state space of the linear system.

Section 2 of this paper contains a review of the transformation theory for taking nonlinear systems to linear systems. In section 3 we present standard definitions and show how to construct Lyapunov functions using the transformation. Our main theory concerning the robustness of the design technique using transformations is proved in section 4. Since the transformation theory exists for time-varying nonlinear systems (see [5-6]; Hunt, L. R., Su, R., and Meyer, G. Time Varying Systems, article in progress), we mention how to extend our results to these systems in the last section.

2. Transformations

We are interested in necessary and sufficient conditions to transform the nonlinear system (1) to the canonical form (2). Historically, theorems in this direction can be found in references [2] and [3], and in the work of Krener [7], Brockett [8], and Jakubczyk and Respondek [9]. Much of the material presented here is discussed in references [10-12].

We begin by introducing the Kronecker indices and Brunovsky canonical forms. Suppose we examine the linear system

$$\dot{y} = Ay + Bv \quad (3)$$

where $y \in \mathbb{R}^n$, $v \in \mathbb{R}^m$, and A and B are appropriately sized matrices. This system is assumed controllable, that is, the span of $\{B, AB, \dots, A^{n-1}B\}$ is n dimensional. We set

$$r_0 = \text{rank } B$$

$$r_j = \text{rank}(B, AB, \dots, A^j B) - \text{rank}(B, AB, \dots, A^{j-1} B),$$

for $1 \leq j \leq n-1$, and have that $0 \leq r_j \leq m$ for $0 \leq j \leq n-1$ and $\sum_{j=0}^{n-1} r_j = n$. We define the Kronecker index κ_j as the number of $r_j \geq 1$ and remark that $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_m$ and $\sum_{j=1}^m \kappa_j = n$.

It is well known that system (3) is equivalent to our linear system in Brunovsky canonical form (2) with a new control v , such that

$$\dot{y} = A_0 y + B_0 v$$

Here A_0 equals

$$\begin{bmatrix} A_1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & A_2 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & 0 & A_m \end{bmatrix}$$

with

$$A_i = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 1 \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix}}_{\kappa_i} \Bigg\} \kappa_i$$

and B_0 equals

$$\begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_m \end{bmatrix}$$

where B_i is the $\kappa_i \times m$ matrix whose only nonzero element is a 1 in the last row and i^{th} column.

We define the type of transformations that are of interest to us. A C^∞ transformation $T = (T_1, T_2, \dots, T_{n+m})$ maps an open set $W \times \mathbb{R}^m$ in \mathbb{R}^{n+m} ($(x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m) = (x, u)$ space) onto an open set in \mathbb{R}^{n+m} ($(y_1, y_2, \dots, y_n, v_1, v_2, \dots, v_m) = (y, v)$ space) containing the origin, where W is an open set in \mathbb{R}^n containing the origin, so that the following properties hold:

i) $T(0) = 0$;

ii) T_1, T_2, \dots, T_n are functions of x only

and have a nonsingular Jacobian matrix in W ;

iii) $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ are functions of (x, u) and for fixed x in W , the $m \times m$ Jacobian matrix of $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ with respect to u is nonsingular;

iv) $T_1 = y_1, T_2 = y_2, \dots, T_n = y_n$ are the state variables and $T_{n+1} = v_1, T_{n+2} = v_2, \dots, T_{n+m} = v_m$ are the controls for the linear time-invariant system (2); in other words, T maps system (1) to system (2); and

v) $T = (T_1, T_2, \dots, T_{n+m})$ is a one-to-one map of $W \times \mathbb{R}^m$ onto an open set containing the origin in (y, v) space.

If we fix the controls $T_{n+1}, T_{n+2}, \dots, T_{n+m}$, then $(u_1, u_2, \dots, u_m) = u$ is fixed, the transformation T is restricted to W , and the only coordinate functions involved are T_1, T_2, \dots, T_n .

We need several additional definitions before stating the theorem on conditions for the existence of transformations from

reference [12]. We mention that a method for constructing a transformation is also available in reference [12].

If f and g are C^∞ vector fields on \mathbb{R}^n , the Lie bracket of f and g is

$$[f, g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g$$

where $\partial g / \partial x$ and $\partial f / \partial x$ are Jacobian matrices. We set

$$(\text{ad}^0 f, g) = \bar{g}$$

$$(\text{ad}^1 f, g) = [f, g]$$

$$(\text{ad}^2 f, g) = [f, [f, g]]$$

.

.

.

$$(\text{ad}^k f, g) = (f, (\text{ad}^{k-1} f, g))$$

A collection of C^∞ vector fields h_1, h_2, \dots, h_r on \mathbb{R}^n is involutive if there exist C^∞ functions γ_{ijk} with

$$[h_i, h_j](x) = \sum_{k=1}^r \gamma_{ijk}(x) h_k(x), \quad 1 \leq i, j \leq r, i \neq j$$

We return to system (1) and define the sets

$$C = \{g_1, [f, g_1], \dots, (\text{ad}^{c_1-1} f, g_1),$$

$$g_2, [f, g_2], \dots, (\text{ad}^{c_2-1} f, g_2),$$

$$\dots, g_m, [f, g_m], \dots, (\text{ad}^{c_m-1} f, g_m)\}$$

$$C_j = \{g_1, [f, g_1], \dots, (\text{ad}^{k_j-2} f, g_1), \\ g_2, [f, g_2], \dots, (\text{ad}^{k_j-2} f, g_2), \\ \dots, g_m, [f, g_m], \dots, (\text{ad}^{k_j-2} f, g_m)\} \text{ for } j = 1, 2, \dots, m.$$

Theorem 2.1

There exists a transformation with properties i) through v) if and only if on W

- 1) the set C spans an n dimensional space,
- 2) each set C_j is involutive for $j = 1, 2, \dots, m$, and
- 3) the span of C_j equals the span of $C_j \cap C$ for $j = 1, 2, \dots, m$.

Theorem 2.1 as stated in reference [12] is local in nature (it holds for some neighborhood of the origin in \mathbb{R}^n), but for the sake of simplicity in notation we have taken this neighborhood as our set W . Global transformation results can be found in reference [11].

There is also a version of this theorem (Hunt, L. R., Su, R., and Meyer, G. Time Varying Systems, article in progress) for time-varying systems which we discuss later.

Since the main topic of this article is the preservation of asymptotic stability under small perturbations of transformable systems, we turn our discussion to stability and Lyapunov functions.

3. Stability

Let \bar{x} be an equilibrium state for the system of differential equations $\dot{x} = h(x)$, where h is C^1 in some set W of the state space which contains \bar{x} .

Definition 3.1

The point \bar{x} is a stable equilibrium if for every neighborhood U of \bar{x} in \mathbb{R}^n there is a neighborhood U_1 of \bar{x} in U such that every solution $x(t)$ with $x(0)$ in U_1 is defined and in U for all $t > 0$. If U_1 can be chosen so that in addition,

$$\lim_{t \rightarrow \infty} x(t) = \bar{x},$$

then \bar{x} is asymptotically stable. Given a compact set K containing \bar{x} in its interior, \bar{x} is called asymptotically stable on K if it is stable and every solution starting in K converges to \bar{x} .

Definition 3.2

Let $V:U \rightarrow \mathbb{R}$ be a continuous function defined on a neighborhood U of \bar{x} , differentiable on $U - \{\bar{x}\}$, such that

- (a) $V(\bar{x}) = 0$ and $V(x) > 0$ if $x \neq \bar{x}$
- (b) $\dot{V} < 0$ in $U - \{\bar{x}\}$.

Then V is a strict Lyapunov function for \bar{x} .

Suppose we take a linear feedback control and apply it to system (2) to stabilize (asymptotically) the system about the origin. With the control substituted we have a linear system of differential equations

$$\dot{y} = C_0 y$$

where the eigenvalues of C_0 have negative real parts. Choosing a negative definite matrix Q , the equation

$$C_0^* P + P C_0 = -Q \quad (* \text{ denotes transpose})$$

yields a unique positive definite solution P , and

$$V(y) = y^* P y$$

is a strict Lyapunov function (see ref. [13], p. 51). We remark that the level surfaces of this Lyapunov function shrink to the origin.

Now $y_1 = T_1, y_2 = T_2, \dots, y_n = T_n$ are functions of x and thus V depends on x . Computing, we find

$$\dot{V} = \sum_{i=1}^n \frac{\partial V}{\partial x_i} \dot{x}_i = \sum_{i=1}^n \left(\sum_{j=1}^n \frac{\partial V}{\partial T_j} \frac{\partial T_j}{\partial x_i} \right) \dot{x}_i$$

$$\dot{V} = \sum_{j=1}^n \frac{\partial V}{\partial T_j} \dot{T}_j = \sum_{j=1}^n \frac{\partial V}{\partial y_j} \dot{y}_j$$

which we know to be negative away from the origin in y space. Hence, V is a strict Lyapunov function for the origin in x -space, the system being (1) with u corresponding to our linear feedback control. We now illustrate how $u = (u_1, u_2, \dots, u_m)$ is computed.

Let $\sigma_1 = \kappa_1$, $\sigma_2 = \kappa_1 + \kappa_2$, \dots , $\sigma_m = \kappa_1 + \kappa_2 + \dots + \kappa_m = n$. From reference [12] we know that u_1, u_2, \dots, u_m and v_1, v_2, \dots, v_m are linearly related by the equations (here (\cdot, \cdot) denotes the duality between one forms and vector fields)

$$\begin{aligned} (dT_{\sigma_1}, f) + \sum_{i=1}^m u_i (dT_{\sigma_1}, g_i) &= v_1 \\ (dT_{\sigma_2}, f) + \sum_{i=1}^m u_i (dT_{\sigma_2}, g_i) &= v_2 \\ &\vdots \\ (dT_n, f) + \sum_{i=1}^m u_i (dT_n, g_i) &= v_m \end{aligned} \quad (4)$$

and the coefficient matrix for u_i is nonsingular under the conditions of Theorem 2.1. Knowing our transformation and the feedback controls v_1, v_2, \dots, v_m , it is easy to compute u_1, u_2, \dots, u_m .

Using Lyapunov's theorem and substituting u_1, u_2, \dots, u_m into system (1), the origin is asymptotically stable. Moreover, if K is a compact subset of W containing the origin and ∂K (the boundary of K) is a level set of $V[y(x)]$, then the origin is asymptotically stable on K . This follows because T (and T^{-1}) map level sets to corresponding level sets, trajectories to corresponding trajectories, and the origin to the corresponding origin. We state this as a theorem.

Theorem 3.3

Suppose we have a transformation T mapping system (1) to system (2) with properties i) through v) holding on an open set W containing the origin in x -space and we use linear feedback in y -space to asymptotically stabilize the linear system (eigenvalues having negative real parts). Any strict Lyapunov function $V(y)$ for the linear system is a strict Lyapunov function $V[y(x)]$ for the nonlinear system with the controls corresponding to those of linear feedback. Moreover, this nonlinear system has the origin as an asymptotically stable equilibrium point on any compact set K whose boundary is a level set of $V[y(x)]$ and with $K \subset W$.

Of course, by taking different Lyapunov functions we can find other sets K and possibly add to the set of points for which solutions starting at these points tend to the origin.

It is appropriate to illustrate this method of constructing Lyapunov functions by an example.

Example 3.4

We take the nonlinear system on R^2

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \sin x_2 \\ 0 \end{bmatrix} + u \begin{bmatrix} 0 \\ 1 \end{bmatrix} = f + ug$$

with $W = \{(x_1, x_2): -\pi/2 < x_2 < \pi/2\}$, an open set containing the origin in R^2 .

It is shown in reference [14] that the hypotheses of Theorem 2.1 are satisfied and that one transformation on W is

$$T_1 = x_1$$

$$T_2 = \sin x_2$$

$$T_3 = (\cos x_2)u$$

The canonical linear system is

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + v \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and $v = -2y_1 - 2y_2$ places the eigenvalues at $-1-i$ and $-1+i$.
We obtain

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = C_0 y$$

With

$$Q = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

we solve $C_0^* P + P C_0 = Q$ to find

$$P = \begin{bmatrix} \frac{5}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{8} \end{bmatrix}$$

Thus, $V(y) = y^* P y = \frac{5}{4} y_1^2 + \frac{1}{2} y_1 y_2 + \frac{3}{8} y_2^2$ is a Lyapunov function for $\dot{y} = C_0 y$ at 0.

Equations (4) in this case become

$$(dT_2, f) + u (dT_2, g) = v \quad \text{or}$$

$$(\cos x_2)u = -2y_1 - 2y_2 = -2x_1 - 2 \sin x_2$$

Solving for u we have

$$u = -\frac{2x_1}{\cos x_2} - 2 \tan x_2$$

which takes our original nonlinear system to

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \sin x_2 \\ -\frac{2x_1}{\cos x_2} - 2 \tan x_2 \end{bmatrix}$$

The function

$$V[y(x)] = \frac{5}{4} x_1^2 + \frac{1}{2} x_1 \sin x_2 + \frac{3}{8} (\sin^2 x_2)$$

is a strict Lyapunov function for the origin. We have derived asymptotic stability for any compact set K with a boundary which is a level set of $V[y(x)]$ and which is contained in W .

We check the eigenvalues of the linearization of the nonlinear system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \sin x_2 \\ -\frac{2x_1}{\cos x_2} - 2 \tan x_2 \end{bmatrix}$$

at the origin. The Jacobian matrix of the right-hand side evaluated at the origin is

$$\begin{bmatrix} 0 & 1 \\ -2 & -2 \end{bmatrix}$$

Hence, the eigenvalues are exactly those of the linear system after linear feedback controls. We show later that these eigenvalues are preserved in general, which is an important step in our robustness theory.

4. Robustness

We will now prove that the method of design using transformations of nonlinear systems to linear systems is robust. First, we need a topology on the set of systems of the form $\dot{x} = h(x)$, that is, a topology on the set of vector fields.

Let $\mathcal{U}(K)$ be the set of \mathcal{C}^1 vector fields on a compact set $K \subset \mathbb{R}^n$. We define the \mathcal{C}^1 -norm $\|h\|_1$ of a vector field $h \in \mathcal{U}(K)$ to be the least upper bound of

$$\|h(x)\|, \quad \|Dh(x)\|$$

for $x \in K$. Here $\|\cdot\|$ is the usual norm on \mathbb{R}^n , Dh is the Jacobian matrix of h (we also denote this by $\partial h / \partial x$) and $\|h\|$ is the usual operator norm on the set of $n \times n$ matrices $L(\mathbb{R}^n)$ (see ref. [15], page 82); that is, for $S \in L(\mathbb{R}^n)$, $\|S\| = \max\{\|Sx\| : \|x\| \leq 1\}$. Two vector fields in $V(K)$ are close in this topology if and only if their corresponding coordinates and the first partial derivatives of these coordinates are close in the usual topology of uniform convergence on K .

For an open set $W \subset \mathbb{R}^n$ and $h \in C^1(W)$, the C^1 vector fields on W , we can take the norm $\|h\|_1$ defined in the same way as above with the understanding that $\|h\|_1 = \infty$ can occur.

Next we examine the linearization of the nonlinear system (1) about 0

$$\dot{x}(t) = Fx + \sum_{i=1}^m u_i(t)g_i^0 \quad (5)$$

where F is the Jacobian matrix of f at 0, and where g_i^0 is the constant part of g_i , $i = 1, 2, \dots, m$, at least one of which is nonvanishing by assumption. We rewrite equations (4) in the vector notation

$$u = H_1^{-1}(x)v + H_2(x)$$

Linear feedback $v = Hy$ yields

$$u = H_1^{-1}(x)Hy + H_2(x)$$

Since $f(0) = 0$, the linear part of $H_2(x)$ is of the form $\tilde{Q}x$.

Ignoring $\tilde{Q}x$, since $T(0) = 0$, the only terms of

$y = (y_1, y_2, \dots, y_n) = (T_1, T_2, \dots, T_n)$ that have an effect on the eigenvalues at 0 in system (5) are those contributed by the linear terms of T_1, T_2, \dots, T_n in the x_1, x_2, \dots, x_n variables.

We are now able to prove the following lemma.

Lemma 4.1

If linear feedback is applied to the linear system (2), then the resulting eigenvalues are the same as those of the linearization (5) of the nonlinear system (1) at the origin with u_1, u_2, \dots, u_m substituted.

Proof

It is apparent from the discussion preceding the statement of this theorem that we have reduced its proof to that of a linear system where the transformations involve linear coordinate changes and feedback. We take the linear system

$$\dot{x} = Fx + Gu$$

and let $z = C^{-1}x$ and $u = H_1^{-1}(0)w + \tilde{Q}x = D\tilde{w} + \tilde{Q}x$. Our linear system in z coordinates with control vector w becomes

$$\dot{z} = C^{-1}(F + G\tilde{Q})Cz + C^{-1}GD\tilde{w}$$

Setting $w = Hz$, which is a feedback on z -space, we obtain

$$\dot{z} = [C^{-1}(F + G\tilde{Q})C + C^{-1}GDH]z$$

Plugging $u = Qx + DHCx^{-1}$ into $Fx + Gu$, we have

$$\dot{x} = [(F + G\tilde{Q}) + GDHC^{-1}]x$$

Since

$$\det.(\lambda I - [C^{-1}(F + G\tilde{Q})C + C^{-1}GDH]) = \det.(\lambda I - [(F + G\tilde{Q}) + GDHC^{-1}])$$

our theorem is proved. \square

The statement in the lemma is not surprising when we realize that since $f(0) = 0$,

$$(ad^{k_{j-1}}g_j)(0) = F^{k_{j-1}}g_j^0$$

for all k , where F and g_i^0 are as in equation (5). Concerning eigenvalues at the origin, the effect of the transformation T on the linearization of system (1) is the same as that of a transformation that takes system (5) to Brunovsky canonical form.

We take the following results from Chapter 16 of reference [15].

Theorem 4.2

Let h be a C^1 vector field on $W \subset \mathbb{R}^n$ and $\bar{x} \in W$ be an equilibrium of $\dot{x} = h(x)$, such that $Dh(\bar{x})$ is invertible. Then there exists a neighborhood $U \subset W$ of \bar{x} and a neighborhood $\mathcal{N} \subset \mathcal{U}(W)$ of h such that for any $\tilde{h} \in \mathcal{N}$ there is a unique equilibrium $\bar{z} \in U$ of $\dot{z} = \tilde{h}(z)$. For any $\epsilon > 0$ we can choose \mathcal{N} so that $|\bar{z} - \bar{x}| < \epsilon$.

Theorem 4.3

Suppose that \bar{x} is an equilibrium point for $\dot{x} = h(x)$ and $Dh(\bar{x})$ has eigenvalues with negative real parts. Then in Theorem 4.2 \mathcal{N} and U can be chosen so that if $\tilde{h} \in \mathcal{N}$, the unique equilibrium $\bar{z} \in U$ of $\dot{z} = \tilde{h}(z)$ satisfies $Dh(\bar{z})$ has eigenvalues with negative real parts.

The following well-known result is found in Chapter 9 of reference [15].

Theorem 4.4

Let $\bar{x} \in W$ be an equilibrium point for $\dot{x} = h(x)$ with $Dh(\bar{x})$ having negative real part eigenvalues. Then there is a neighborhood $U' \subset W$ of \bar{x} such that all solutions of $\dot{x} = h(x)$ starting in U' converge to \bar{x} ; thus, \bar{x} is asymptotically stable.

This theorem is a shortened version of the first theorem in Chapter 9 (see ref. [15]). If every eigenvalue of $Dh(\bar{x})$ has its real part less than $-c$, $c > 0$ it follows from the proof of that theorem that the set U' depends on c and continuously on $h(x)$.

It is clear that if the hypotheses of Theorems 4.2 and 4.3 are satisfied, we can choose the set \mathcal{N} in Theorem 4.3 so that there exists a constant $c > 0$ so that the eigenvalues of $Dh(\bar{z})$ have real parts less than $-c$ for all $\tilde{h} \in \mathcal{N}$. Since we are working with the C^1 topology on the set of systems, the way in which we choose the sets U' in Theorem 4.4 depends continuously on elements in \mathcal{N} .

If the conditions of Theorem 4.3 are met, then we can choose \mathcal{N} and an open neighborhood $O \subset W$ of \bar{x} so that the following properties hold:

- (a) for any $\bar{h} \in \Pi$ there is a unique equilibrium $\bar{z} \in O$ of $\dot{z} = \bar{h}(z)$ and $D\bar{h}(\bar{z})$ has eigenvalues with negative real parts, and
- (b) for any $\bar{h} \in \Pi$ all solutions of $\dot{z} = \bar{h}(z)$ starting in O converge to \bar{z} .

In our application, h will be $f + \sum_{i=1}^m u_i g_i$ where $u_i, 1 \leq i \leq m$, are controls corresponding to a linear feedback giving eigenvalues with negative real parts in the linear system (2), and $\bar{x} = 0$. Of course, we assume that the nonlinear system (1) is transformable for $x \in W$ to the linear system, as in Theorem 2.1. From section 3, Theorem 3.3, we have a strict Lyapunov function V for O and the system

$$\dot{x} = f + \sum_{i=1}^m u_i g_i$$

and we take K to be a compact set contained in W whose boundary is a level set of $V(x)$. We now state our main result.

Theorem 4.5

With controls u_1, u_2, \dots, u_m and compact set K as just mentioned, there exists an open neighborhood Π of $f + \sum_{i=1}^m u_i g_i$ in $\psi(K)$ such that

- (a') for any $\bar{h} \in \Pi$ there is a unique equilibrium point $\bar{z} \in K$ of $\dot{z} = \bar{h}(z)$, and
- (b') for any $\bar{h} \in \Pi$ all solutions of $\dot{z} = \bar{h}(z)$ starting in K converge to \bar{z} ; that is, \bar{z} is asymptotically stable on K .

Proof

Lemma 4.1 implies that the system $\dot{x} = f + \sum_{i=1}^m u_i g_i$, with u_i corresponding to our asymptotically stabilizing linear feedback in the linear system, has eigenvalues with negative real parts in its linearization at O . From a previous discussion we know that there is a neighborhood Π' in $\psi(K)$ [$\psi(W)$ in previous results can be replaced by $\psi(K)$] of $f + \sum_{i=1}^m u_i g_i$ and an open neighborhood O in \mathbb{R}^n of O with our earlier conditions (a) and (b) holding for Π' on O .

Choose a compact annular region $\partial K \subset K$, whose outer boundary is the level set ∂K (the boundary of K), and whose inner boundary

is another level set of V contained in O and distinct from O . We recall here that our level sets of V do shrink to the origin. Because V is a strict Lyapunov function of $\dot{x} = f + \sum_{i=1}^m u_i g_i$, we have $\dot{V} = \langle dV, f + \sum_{i=1}^m u_i g_i \rangle < 0$. Since we are working in the e^1 topology on $V(K)$, we can choose $\pi \subset \pi'$ so that $\dot{V} = \langle dV, \tilde{h} \rangle < 0$ on Ω for all $\tilde{h} \in \pi$ and \tilde{z} is the unique equilibrium point for \tilde{h} in K . The condition that $\langle dV, \tilde{h} \rangle < 0$ implies that the solution curves of $\dot{z} = \tilde{h}(z)$ intersect the level sets of V in Ω transversally (the tangents to the solution curves are not tangent to the level sets at points of intersection) and must be moving from outside to inside, since those for the nearby system $\dot{x} = f + \sum_{i=1}^m u_i g_i$ are transversing in that direction.

Hence, if we start with any point in K with a solution of $\dot{z} = \tilde{h}(z)$, we can reach the set O . Then condition (B) guarantees that we converge to \tilde{z} , and statements (a') and (b') are proved. \square

In practice, suppose the plant is a system $\dot{z} = \tilde{h}(z)$, where $\tilde{h} \in \pi$ is driven by the controls u_1, u_2, \dots, u_m . Then Theorem 4.5 proves the robustness of our nonlinear design technique.

We conclude this section with a remark concerning systems which depend on parameter values. Let system (1) depend on a parameter θ , and assume for $\theta = \theta_0$ we have a transformable system. Suppose we design with the parameter value θ_0 and transformation $T(\theta_0)$. We construct a Lyapunov function $V(y)$ as before and compose with $T(\theta_0)$. The sensitivity of the asymptotic stability of

$$\dot{x}(t) = f[x(t), \theta] + \sum_{i=1}^m u_i(t) g_i[x(t), \theta]$$

can be viewed by examining the θ derivative of

$$\sum_{k=1}^n \left[\sum_{j=1}^n \frac{\partial V}{\partial T_j(\theta_0)} \frac{\partial T_j(\theta_0)}{\partial x_k} \right] \left[f_k(\theta) + \sum_{i=1}^m u_i(t) g_{ik}(\theta) \right]$$

where $g_{ik}(\theta)$ is the k^{th} component of $g_i[x(t), \theta]$ and similarly for f_k .

5. Time-Varying Systems

Consider the time-varying nonlinear system

$$\dot{x}(t) = f(x,t) + \sum_{i=1}^m u_i(t) g_i(x,t) \quad (6)$$

with f, g_1, g_2, \dots, g_m being vector fields on \mathbb{R}^n which are C^∞ in all arguments. Also assume that $f(0,t) = 0$ for all t . We want a transformation $T = (T_1, T_2, \dots, T_{n+m})$ that maps an open set in $\mathbb{R}^{n+m+1} [(x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m, t) \text{ space}]$ containing the origin in \mathbb{R}^{n+m} onto an open set in $\mathbb{R}^{n+m} [(y_1, y_2, \dots, y_n, v_1, v_2, \dots, v_m) = (T_1, T_2, \dots, T_{n+m}) \text{ space}]$ containing the origin and takes the time-varying system (6) to the time-invariant linear system (2). We also want conditions analogous to i) through v) in section 2 to hold, except here we must remember that T is a function of time t . Thus, in each of i) through v) we add the phrase "for every t ."

Our Lie derivative is now replaced by a time-varying Lie derivative. Suppose f and g are C^∞ time-varying vector fields. Then

$$(\tau^0 f, g) = g$$

$$(\tau^1 f, g) = (\text{ad}^1 f, g) + \frac{dg}{dt}$$

.

.

.

$$(\tau^k f, g) = [\tau^1 f, (\tau^{k-1} f, g)]$$

In this case we define new sets $C(t)$ and $C_j(t)$, $j = 1, 2, \dots, m$, by replacing $(\text{ad}^k f, g)$ in our definitions for C and C_j by $(\tau^k f, g)$. The following result is found in Hunt, L. R., Su, R., and Meyer, G. (Time Varying Systems, article in progress).

Theorem 5.1

There exists a transformation T satisfying our new conditions i) through v) if and only if on some neighborhood W of the origin in R^n

- 1) the set $C(t)$ spans an n dimensional space for each t ,
- 2) each set $C_j(t)$ is involutive for $j = 1, 2, \dots, m$ and each fixed t , and
- 3) the span of $C_j(t)$ equals the span of $C_j(t) \cap C(t)$ for $j = 1, 2, \dots, m$ and each t .

It is shown in Hunt, L. R. et al. (article in progress) that the transformation T is constructed, with t behaving as a parameter.

We apply linear feedback to the time-invariant linear system (2) to asymptotically stabilize (with real part negative eigenvalues) this system. We substitute the corresponding controls into equation (6) using equations from Hunt, L. R. et al. (article in progress) similar to those in equation (4).

Let $V(y)$ be a Lyapunov function for the linear system (2) as in section 3 and compose this with the transformation T to obtain $V[y(x,t)]$. Since $V[y(x,t)] < 0$, a solution in x_1, x_2, \dots, x_n, t space is passing through the level sets (the inverse images of the level sets of $V(y)$ under T) transversally and from the outside to the inside. Suppose there are positive definite functions $V_1(x)$ and $V_2(x)$ ($V_1(x)$ has continuous first partial derivatives, $V_1(0) = 0$, $V_1(x) \geq 0$, $i = 1, 2$) on some compact set K in x_1, x_2, \dots, x_n space with the origin in its interior. Suppose also that the transformation T applies to K for all t and $V_1(x) \leq V[y(x,t)] \leq V_2(x)$ for $x \in K$ and all t . By Theorem 4.2 from reference [13] we know the origin is uniformly asymptotically stable for the system (6) with the controls u_1, u_2, \dots, u_m substituted. In fact, any solution starting at a point in the region bounded by a level set of $V[y(x,t)]$ with $x \in K$ for all t must tend to the origin in x -space.

There are many results concerning stability of time-varying systems that we could mention. However, we wish to make the following remark concerning robustness of time-varying systems. Practice dictates that a theory is to be applied in finite time.

Let us assume that the transformation T applies to an annular region A about the t -axis whose outer boundary and inner boundary are nontrivial level sets of $V[y(x,t)]$. Suppose we restrict this region to a finite time interval, say $[0, t_0]$, $t_0 > 0$. If a system $\dot{x} = h(x,t)$ is sufficiently close to system (6) (with u_1, u_2, \dots, u_m applied) in the obvious C^1 topology for vector fields on our truncated annulus A , then the solutions of $\dot{x} = h(x,t)$ starting in this set travel transversally through the level sets (it intersects) contained in A moving from outside to inside. This parallels the related results for time-invariant systems in section 4.

In conclusion, we have shown in this paper that the process of stabilizing a transformable nonlinear system by stabilizing the linear system to which it maps is robust. If the mathematical model and the actual nonlinear plant are sufficiently close, our design scheme using transformations is valid, and Lyapunov functions can often be found.

The design technique discussed has been applied to several aircraft of increasing complexity. The completely automatic flight control system was first tested on a DHC-6. The reference trajectory used in the flight test exercised a substantial part of the operational envelope of the aircraft. Despite disturbances and variations in plant dynamics, the system performed well (see ref. [16]). Next the technique was applied to the Augmentor Wing Jet STOL Research aircraft and the successful flight tests are reported in reference [3]. Methods for providing pilot inputs to this design were examined in reference [17]. Application of this design scheme to the control of an A-7 for carrier landing and testing in manned simulation is reported in references [18] and [19]. The design methodology is currently being applied to the UH-1H helicopter, again with the substantial portion of the operational envelope of this aircraft being used.

References

- [1] Brunovsky, P. 1970. A classification of linear controllable systems. *Kibernetika (Praga)* 6:173-188.
- [2] Meyer, G., and Ciccolani, L. 1975. A Formal Structure for Advanced Flight Control Systems. NASA TN D-7940.

- [3] Meyer, G., and Cicolani, L. 1980. Applications of nonlinear system inverses to automatic flight control design--system concepts and flight evaluations. AGARDograph 251 on theory and applications of optimal control in aerospace systems. P. Kant, ed., NATO.
- [4] Meyer, G. 1981. The design of exact nonlinear model followers. In 1981 Joint Automatic Control Conference, FA-3A.
- [5] Hunt, L. R., and Su, R. 1981. Linear equivalents of linear time-varying systems. In 1981 International Symposium on Mathematical Theory of Networks and Systems, pp. 119-123.
- [6] Hunt, L. R., and Su, R. 1981. Control of nonlinear time-varying systems. In 1981 IEEE Conference on Decision and Control, pp. 558-563.
- [7] Krener, A. J. 1973. On the equivalence of control systems and linearization of nonlinear systems. SIAM J. Control 11, pp. 670-676.
- [8] Brockett, R. W. 1978. Feedback invariants for nonlinear systems. In IFAC Congress, Helsinki.
- [9] Jakubczyk, B., and Respondek, W. 1980. On linearization of control systems. Bull. Acad. Polon. Sci., Ser. Sci. Math. Astronom. Phys. 28:517-522.
- [10] Su, R. 1982. On the linear equivalents of nonlinear systems. To appear in Systems and Control Letters, Vol 2, No. 1.
- [11] Hunt, L. R., Su, R., and Meyer, G. 1982. Global transformations of nonlinear systems. To appear in IEEE Trans. on Autom. Control, Vol. 27.
- [12] Hunt, L. R., Su, R., and Meyer, G. 1982. Multi-input nonlinear systems. To appear in Differential Geometric Control Theory Conference, Birkhäuser Boston, Cambridge, Mass.
- [13] Williams, J. L. 1970. Stability Theory of Dynamical Systems. John Wiley and Sons, Inc: New York.
- [14] Hunt, L. R., and Su, R. 1981. The Poincaré lemma and transformations of nonlinear systems. In 1981 International Symposium on Mathematical Theory of Networks and Systems, pp. 111-118.
- [15] Hirsch, M. W., and Smale, S. 1974. Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press: New York.
- [16] Wehrend, W. R., Jr., and Meyer, G. 1980. Flight Tests of the Total Automatic Flight Control System (TAF COS) Concept on a DHC-6 Twin Otter Aircraft. NASA TP-1513.
- [17] Wehrend, W. R., Jr. 1979. Pilot Control Through the TAF COS Automatic Flight Control System. NASA TM-81152.
- [18] Smith, G. A., and Meyer, G. 1979. Total aircraft flight control system balanced open- and closed-loop control with dynamic trimmaps. In 3rd Avionics Conference, Dallas, TX.
- [19] Smith, G. A., and Meyer, G. 1980. Application of the Concept of Dynamic Trim Control to Automatic Landing of Carrier Aircraft. NASA TP-1512.

Global Transformations of Nonlinear Systems

L. R. HUNT, MEMBER, IEEE, RENJENG SU, AND GEORGE MEYER, MEMBER, IEEE

Abstract—Recent results have established necessary and sufficient conditions for a nonlinear system of the form

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)),$$

with $f(0) = 0$, to be locally equivalent in a neighborhood of the origin in \mathbb{R}^n to a controllable linear system. We combine these results with several versions of the global inverse function theorem to prove sufficient conditions for the transformation of a nonlinear system to a linear system. In doing so we introduce a technique for constructing a transformation under the assumptions that $\{g, [f, g], \dots, (ad^{n-1}f, g)\}$ span an n -dimensional space and that $\{g, [f, g], \dots, (ad^{n-2}f, g)\}$ is an involutive set.

I. INTRODUCTION

WE CONSIDER nonlinear systems of the form

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)) \quad (1)$$

where f and g are \mathcal{C}^∞ vector fields on an open set U in \mathbb{R}^n containing the origin and $f(0) = 0$. The problem of interest to us is finding sufficient conditions on U , f , and g so that there exists a \mathcal{C}^∞ transformation $T = (T_1, T_2, \dots, T_{n+1})$ from an open set in \mathbb{R}^{n+1} to an open set in \mathbb{R}^{n+1} with the following properties:

- 1) $T(0) = 0$;
- 2) T_1, T_2, \dots, T_n are functions of x_1, x_2, \dots, x_n only and T maps U in \mathbb{R}^n into T_1, T_2, \dots, T_n space with a nonsingular Jacobian matrix;
- 3) T_{n+1} is a function of x_1, x_2, \dots, x_n, u which can be inverted as a function of u and where $(x_1, x_2, \dots, x_n) \in U$;
- 4) T_1, T_2, \dots, T_{n+1} satisfy

$$\begin{aligned} \dot{T}_1 &= T_2 \\ \dot{T}_2 &= T_3 \\ &\vdots \\ \dot{T}_n &= T_{n+1} \end{aligned} \quad (2)$$

Manuscript received June 11, 1982; revised March 22, 1982. Paper recommended by A. N. Michel, Past Chairman of the Stability, Nonlinear, and Distributed Systems Committee. The work of L. R. Hunt was supported by Ames Research Center, NASA, under the IPA Program and the Joint Services Electronics Program, Texas Tech University, under ONR Contract N00014-76-C-1136. The work of R. Su was supported by Ames Research Center, NASA, and the National Research Council.

L. R. Hunt was on leave at Ames Research Center, NASA, Moffett Field, CA 94035. He is with Texas Tech University, Lubbock, TX 79409.

R. Su was with Ames Research Center, NASA, Moffett Field, CA. He is now with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

G. Meyer is with Ames Research Center, NASA, Moffett Field, CA 94035.

5) T is one-to-one (with (T_1, T_2, \dots, T_n) being one-to-one on U).

That is, when can we map a nonlinear system in x_1, x_2, \dots, x_n, u space with $(x_1, x_2, \dots, x_n) \in U$ to a linear system in T_1, T_2, \dots, T_{n+1} space taking the system (1) to the system (2)? In (2) we think of T_{n+1} as the control, and our linear system is in integrator form on n -dimensional space.

Krener [1] solved the problem of which nonlinear systems can be transformed to linear systems, using a change of variable without feedback. Brockett [2] gave sufficient conditions for a real-analytic nonlinear system with equilibrium point at the origin to be locally equivalent (using coordinate changes and feedback) to a linear system in integrator form. The third author and Cicolani [3], [4] presented sufficient conditions for a nonlinear system (perhaps time varying) in block-triangular form to be transformed to a linear system. These results are presently being applied in the design of automatic flight control systems for aircraft.

The references just mentioned led the second author [5] to prove necessary and sufficient conditions for a local diffeomorphism (mapping origin to origin) to exist which carries system (1) to system (2). The transformations involved contain both feedback and coordinate changes and are more general than those found in [2]. Proofs depend on the solution of an overdetermined system of linear partial differential equations.

We combine these results with global inverse function theorems and techniques in partial differential equations to prove global theorems for transforming a nonlinear system to a linear system. Local results tell us that there is a neighborhood of the origin which is mapped under a transformation, but give us no information about its size.

A brief history of global inverse function theorems is appropriate. Hadamard [6]–[8] proved the following result. A \mathcal{C}^∞ map $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a diffeomorphism onto \mathbb{R}^n if and only if its Jacobian matrix is nonsingular for each $x \in \mathbb{R}^n$ and F is proper (i.e., inverse images of compact sets are compact). Palais [9] gave a proof of this theorem, and certain variations of inverse function results are presented in [10].

The question of the existence of global inverses is treated in the area of systems theory by Wu and Deser [11] and by Kuh and Hajj [12]. With regard to the problem of global observability, we find the papers of Fitts [13], Griffith and Kumar [14], and Fujisawa and Kuh [15]. Applying the results of Fujisawa and Kuh and of Berger and Berger [10], Kou *et al.* [16] presented theorems on the

topic of global observability, using inverse functions. For recent results on global inverse functions, we refer to the papers of Sandburg [17] and Miller [18].

Given vector fields f and g , we denote the Lie bracket of f and g by $[f, g]$, and define inductively

$$(ad^2 f, g) = [f, [f, g]], \dots$$

$$(ad^k f, g) = [f, (ad^{k-1} f, g)].$$

Given system (1), we suppose that the $n \times n$ matrix with columns $g, [f, g], \dots, (ad^{n-1} f, g)$ (which in this paper we call the controllability matrix, even though for nonlinear systems controllability can certainly depend on other Lie brackets than these [19]–[21]) is nonsingular at every point in \mathbb{R}^n , and $g, [f, g], \dots, (ad^{n-2} f, g)$ are involutive on \mathbb{R}^n . We introduce a certain matrix with the property that if it satisfies the ratio condition found in [16], we show the existence of a transformation $T = (T_1, T_2, \dots, T_{n-1})$ from \mathbb{R}^{n-1} to \mathbb{R}^{n-1} with (T_1, T_2, \dots, T_n) mapping \mathbb{R}^n to \mathbb{R}^n (and having a nonsingular Jacobian everywhere on \mathbb{R}^n) such that the system (1) is mapped to the system (2). Under these assumptions the map is one-to-one, and we have global invertibility.

If U is an open subset of \mathbb{R}^n containing the origin, we give sufficient conditions for there to exist a transformation $T = (T_1, T_2, \dots, T_{n-1})$ with (T_1, T_2, \dots, T_n) mapping U one-to-one onto its image.

In all cases, the proof of the existence of our transformations is constructive in nature, which should be useful in future applications.

II. PRELIMINARIES

In this section we discuss some basic definitions, a classical result from differential geometry, and the basic ideas of the theory in [5]. We begin by introducing three kinds of Lie derivatives and a Leibnitz-type relation between them. An enlightening discussion of these Lie derivatives and their applications in system theory is presented in the paper of Hermann and Krener [22]. For basic references we suggest [23]–[25], and [26].

If f and g are \mathcal{C}^∞ vector fields on \mathbb{R}^n (actually on any differentiable manifold M), we define the Lie bracket of f and g

$$[f, g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g$$

where $\partial g / \partial x$ and $\partial f / \partial x$ denote $n \times n$ Jacobian matrices. This Lie bracket $[f, g]$ is also a vector field on \mathbb{R}^n and represents the Lie derivative of one vector field with respect to another. Of course, we can also discuss successive Lie brackets $[f, [f, g]], [g, [f, g]]$, etc. We define

$$(ad^1 f, g) = [f, g]$$

$$(ad^2 f, g) = [f, [f, g]]$$

$$(ad^k f, g) = [f, (ad^{k-1} f, g)].$$

A set of \mathcal{C}^∞ vector fields $\{f_1, f_2, \dots, f_r\}$ on \mathbb{R}^n is called involutive if there exist \mathcal{C}^∞ functions $\gamma_{ij}(x)$ such that

$$[f_i, f_j](x) = \sum_{k=1}^r \gamma_{ijk}(x) f_k(x), \quad 1 \leq i, j \leq r, i \neq j.$$

This definition leads to a version of the famous Frobenius theorem.

Theorem 2.1: Let f_1, f_2, \dots, f_r be an involutive collection of \mathcal{C}^∞ linearly independent vector fields on \mathbb{R}^n . Given any point $x_0 \in \mathbb{R}^n$, there exists a unique maximal r -dimensional \mathcal{C}^∞ submanifold S of \mathbb{R}^n containing x_0 such that the tangent space to S at each $x \in S$ is the space spanned by $f_1(x), f_2(x), \dots, f_r(x)$; that is, S is the unique integral manifold of f_1, f_2, \dots, f_r through x_0 .

We now define the Lie derivative of a function with respect to a vector field, which takes functions to functions. Let f be a \mathcal{C}^∞ vector field and h a \mathcal{C}^∞ function, with gradient dh on \mathbb{R}^n . Then the Lie derivative of h with respect to f is

$$L_f(h) = \langle dh, f \rangle$$

where $\langle \cdot, \cdot \rangle$ denotes the duality between one-forms and vector fields (the cotangent bundle and tangent bundle if one is working on a manifold M). This duality is easily understood if we set

$$\left\langle dx_i, \frac{\partial}{\partial x_j} \right\rangle = \begin{cases} 0, & i \neq j \\ 1, & i = j. \end{cases}$$

If f is a \mathcal{C}^∞ vector field on \mathbb{R}^n and ω is a \mathcal{C}^∞ one-form on \mathbb{R}^n , i.e., $\omega = \omega_1 dx_1 + \omega_2 dx_2 + \dots + \omega_n dx_n$, ω_i being \mathcal{C}^∞ functions, we have the Lie derivative of ω with respect to f

$$L_f(\omega) = \left(\frac{\partial \omega^*}{\partial x} f \right)^* + \omega \frac{\partial f}{\partial x}$$

where $*$ denotes transpose and $\partial \omega^* / \partial x$ and $\partial f / \partial x$ are Jacobian matrices. Such a Lie derivative maps one-forms to one-forms.

The above three kinds of Lie differentiation are related by a Leibnitz-type formula

$$L_f \langle \omega, g \rangle = \langle L_f(\omega), g \rangle + \langle \omega, [f, g] \rangle$$

with f and ω as before and g a \mathcal{C}^∞ vector field. Also, $dL_f(h) = L_f(dh)$ with h a \mathcal{C}^∞ function.

We examine the problem of finding a map with nonsingular Jacobian matrix of the system on \mathbb{R}^n

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)) \quad (1)$$

with $f(0) = 0$ to the linear system

$$\begin{aligned} \dot{T}_1 &= T_2 \\ \dot{T}_2 &= T_3 \\ &\vdots \\ \dot{T}_n &= T_{n+1} \end{aligned} \quad (2)$$

From [5] necessary and sufficient conditions for the local existence of such a map are that

i) the controllability matrix $\langle g, [f, g], \dots, (ad^{n-1}f, g) \rangle$ has rank n in some neighborhood of the origin in \mathbb{R}^n (with variables x_1, x_2, \dots, x_n), and

ii) the set of vector fields $\langle g, [f, g], \dots, (ad^{n-2}f, g) \rangle$ is involutive in some neighborhood of the origin in \mathbb{R}^n .

Note that by i) $g, [f, g], \dots, (ad^{n-2}f, g)$ are linearly independent and by ii) there is an $n-1$ -dimensional integral manifold of this involutive set by the Frobenius theorem.

The question arises as to how the existence of a local (in the neighborhood of the origin) invertible transformation T is shown in [5] under Assumptions i) and ii). Such a map $T = (T_1, T_2, \dots, T_{n-1})$ must satisfy the following partial differential equations:

$$\begin{aligned} \frac{\partial T_i}{\partial x_1} g_1 + \frac{\partial T_i}{\partial x_2} g_2 + \dots + \frac{\partial T_i}{\partial x_n} g_n &= 0, \\ i &= 1, 2, \dots, n-1, \\ \frac{\partial T_i}{\partial x_1} f_1 + \frac{\partial T_i}{\partial x_2} f_2 + \dots + \frac{\partial T_i}{\partial x_n} f_n &= T_{i+1}, \\ i &= 1, 2, \dots, n-1, \\ \frac{\partial T_n}{\partial x_1} (f_1 + ug_1) + \frac{\partial T_n}{\partial x_2} (f_2 + ug_2) + \dots \\ &+ \frac{\partial T_n}{\partial x_n} (f_n + ug_n) = T_{n+1}, \quad (3) \end{aligned}$$

which we write as

$$\begin{aligned} \langle dT_i, g \rangle &= 0, \quad i = 1, 2, \dots, n-1, \\ \langle dT_i, f \rangle &= L_f(T_i) = T_{i+1}, \quad i = 1, 2, \dots, n-1, \\ \langle dT_n, f + ug \rangle &= L_{f+ug}(T_n) = T_{n+1}. \quad (4) \end{aligned}$$

Note that T_1, T_2, \dots, T_n are functions of x_1, x_2, \dots, x_n only and that T_{n+1} is a function of x_1, x_2, \dots, x_n and u .

Now by our Leibnitz formula

$$\begin{aligned} \langle dT_2, g \rangle &= \langle L_f(dT_1), g \rangle \\ &= L_f \langle dT_1, g \rangle - \langle dT_1, [f, g] \rangle \\ &= 0 - \langle dT_1, [f, g] \rangle, \\ \langle dT_3, g \rangle &= \langle L_f(dT_2), g \rangle \\ &= L_f \langle dT_2, g \rangle - \langle dT_2, [f, g] \rangle \\ &= 0 + \langle dT_1, [f, [f, g]] \rangle \\ &= \langle dT_1, (ad^2 f, g) \rangle \end{aligned}$$

and

$$\langle dT_n, g \rangle = \langle L_f(dT_{n-1}), g \rangle = (-1)^{n-1} \langle dT_1, (ad^{n-1}f, g) \rangle.$$

Thus if we know T_1 , then T_2, T_3, \dots, T_{n-1} can be found by Lie differentiation. This fact is also pointed out in [27], but no conditions for the existence of T_1 are given there. Thus (4) becomes

$$\begin{aligned} \langle dT_1, (ad^k f, g) \rangle &= 0, \quad k = 0, 1, \dots, n-2 \\ &\text{with } (ad^0 f, g) = g \\ \langle dT_n, f + ug \rangle &= T_{n+1}. \end{aligned} \quad (5)$$

It is shown in [5] that T_1, T_2, \dots, T_{n-1} have linearly independent gradients if and only if the controllability matrix $\langle g, [f, g], \dots, (ad^{n-1}f, g) \rangle$ has rank n if and only if we have

$$\begin{aligned} \langle dT_1, (ad^k f, g) \rangle &= 0, \quad k = 0, 1, \dots, n-2 \\ \langle dT_1, (ad^{n-1}f, g) \rangle &= 0. \end{aligned} \quad (6)$$

If the last equation in (6) is satisfied, then we can solve for u in the last equation in (5) since $\langle dT_n, g \rangle = (-1)^{n-1} \langle dT_1, (ad^{n-1}f, g) \rangle$ as before. This, together with the fact that T_1, T_2, \dots, T_n are functions of x_1, x_2, \dots, x_n , only turns our problem of finding a global transformation T into one of finding a global solution of (6). With this in mind, we turn to the question of solvability of partial differential equations.

III. PARTIAL DIFFERENTIAL EQUATIONS

We shall solve the overdetermined system of linear partial differential equations (6). For an introductory discussion of this topic, we refer the reader to the book of John [28].

We introduce the parameter s and begin our construction of a C^∞ solution T_1 of the system

$$\begin{aligned} \langle dT_1, (ad^k f, g) \rangle &= 0, \quad k = 0, 1, \dots, n-2 \\ \langle dT_1, (ad^{n-1}f, g) \rangle &= 0. \end{aligned} \quad (6)$$

First we solve for all $s \in \mathbb{R}$ the system

$$\frac{dx}{ds} = (ad^{n-1}f, g)$$

with initial conditions $x(0) = 0$ to find the integral curve of the vector field $(ad^{n-1}f, g)$ through the origin in n space. We remark that we could start with any vector field that is linearly independent of $g, [f, g], \dots, (ad^{n-2}f, g)$.

Next we solve for all $t_1 \in \mathbb{R}$ the system

$$\frac{dx}{dt_1} = (ad^{n-2}f, g)$$

with initial conditions $x(s, 0) = x(s)$. To solve the partial differential equation $\langle dT_1, (ad^{n-2}f, g) \rangle = 0$ we must find a function T_1 so that $\partial T_1 / \partial t_1 = 0$; thus, we have reduced to a system of ordinary differential equations. We then examine for all $t_2 \in \mathbb{R}$ the system

$$\frac{dx}{dt_2} = (ad^{n-3}f, g)$$

with initial conditions $x(s, t_1, 0) = x(s, t_1)$. To solve $\langle dT_1, (ad^{n-3}f, g) \rangle = 0$, we show the existence of a T_1 with $\partial T_1 / \partial t_2 = 0$.

We repeat this argument, with the last step being to consider for all $t_{n-1} \in \mathbb{R}$ the system

$$\frac{dx}{dt_{n-1}} = g$$

having initial conditions $x(s, t_1, \dots, t_{n-2}, 0) = x(s, t_1, \dots, t_{n-2})$.

t_{n-2}). For $\langle dT_1, g \rangle = 0$ we need a function T_1 with $\partial T_1 / \partial t_{n-1} = 0$.

For our solution we choose $T_1 = s$ (actually any C^∞ function of s which vanishes at $(0, 0, \dots, 0)$ and with nonvanishing derivative will work). Since the set $\{g, [f, g], \dots, (ad^{n-2}f, g)\}$ is involutive, for each fixed s the map

$$F = (x_1(s, t_1, \dots, t_{n-1}), x_2(s, t_1, \dots, t_{n-1}), \dots, x_n(s, t_1, \dots, t_{n-1}))$$

defines an integral manifold of this involutive set. Hence $T_1 = s$ does solve the desired equations as a function of s, t_1, \dots, t_{n-1} since it is constant on every such integral manifold and

$$\frac{\partial T_1}{\partial t_1}, \frac{\partial T_1}{\partial t_2}, \dots, \frac{\partial T_1}{\partial t_{n-1}}$$

all vanish.

The important question to answer is when can the map F be inverted in the sense that we solve for s, t_1, \dots, t_{n-1} as functions of x_1, x_2, \dots, x_n ? Before addressing this problem, we show that the last equation in (6)

$$\langle dT_1, (ad^{n-1}f, g) \rangle = 0$$

is solved.

Computing, we have

$$\begin{aligned} dT_1 &= \frac{\partial T_1}{\partial x_1} dx_1 + \frac{\partial T_1}{\partial x_2} dx_2 + \dots + \frac{\partial T_1}{\partial x_n} dx_n \\ &= \frac{\partial T_1}{\partial x} \frac{\partial s}{\partial x_1} dx_1 + \frac{\partial T_1}{\partial s} \frac{\partial s}{\partial x_2} dx_2 + \dots \\ &\quad + \frac{\partial T_1}{\partial s} \frac{\partial s}{\partial x_n} dx_n \\ &= \frac{\partial T_1}{\partial s} ds = ds. \end{aligned}$$

If $\langle dT_1, (ad^{n-1}f, g) \rangle = 0$ since dT_1 and $(ad^{n-1}f, g)$ are both nonzero, we have that $(ad^{n-1}f, g)$ must be tangent to an integral manifold of $\{g, [f, g], \dots, (ad^{n-2}f, g)\}$, contradicting the fact that the controllability matrix has rank n .

Hence, our problem is solved once we know that we can find s, t_1, \dots, t_{n-1} as functions of x_1, x_2, \dots, x_n . This depends on the Jacobian matrix of F

$$\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t_1} & \dots & \frac{\partial x_1}{\partial t_{n-1}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial s} & \frac{\partial x_n}{\partial t_1} & \dots & \frac{\partial x_n}{\partial t_{n-1}} \end{bmatrix} \quad (7)$$

being nonsingular (this is called the noncharacteristic condition in partial differential equations). By design, this matrix is the controllability matrix along the integral curve of $(ad^{n-1}f, g)$ through the origin. Hence, there is an open

neighborhood of the origin on which the map F is invertible.

Thus conditions for showing the existence of a global transformation $T = (T_1, T_2, \dots, T_n)$ can be interpreted in terms of the matrix (7).

Before moving on to global considerations, we examine the Jacobian matrix of (T_1, T_2, \dots, T_n) with respect to s, t_1, \dots, t_{n-1} . Equations (6) imply that

$$\frac{\partial T_1}{\partial t_1}, \frac{\partial T_1}{\partial t_2}, \dots, \frac{\partial T_1}{\partial t_{n-1}}$$

are all zero. Applying the Leibnitz formula once to (6) we find

$$\langle dT_2, (ad^k f, g) \rangle = 0, \quad k = 0, 1, \dots, n-3$$

$$\langle dT_2, (ad^{n-2}f, g) \rangle = 0.$$

Thus, we have

$$\frac{\partial T_2}{\partial t_2}, \frac{\partial T_2}{\partial t_3}, \dots, \frac{\partial T_2}{\partial t_{n-1}}$$

must vanish. Similarly,

$$\langle dT_3, (ad^k f, g) \rangle = 0, \quad k = 0, 1, \dots, n-4$$

$$\langle dT_3, (ad^{n-3}f, g) \rangle = 0$$

imply

$$\frac{\partial T_3}{\partial t_3} = 0, \frac{\partial T_3}{\partial t_4} = 0, \dots, \frac{\partial T_3}{\partial t_{n-1}} = 0.$$

Continuing in this manner, we find that the Jacobian matrix of T_1, T_2, \dots, T_n with respect to s, t_1, \dots, t_{n-1} has all entries above the diagonal zero.

IV. GLOBAL TRANSFORMATION RESULTS

We give sufficient conditions for a global transformation (of the type of interest to us) to exist from a nonlinear system to a linear system.

Our theory depends on the Jacobian matrix (7) of the map F that we constructed in the last section. We refer to this matrix as the *noncharacteristic matrix* because of its application in partial differential equations. Note that this matrix does not depend on the T_1 map, but only on the functions $x_1(s, t_1, \dots, t_{n-1}), x_2(s, t_1, \dots, t_{n-1}), \dots, x_n(s, t_1, \dots, t_{n-1})$.

The first result we need can be found in [16].

Theorem 4.1: Suppose that there is a map $H: \mathbb{R}^n \rightarrow \mathbb{R}^n$ which is differentiable with Jacobian matrix $J(x)$. If there exists a constant $\epsilon > 0$ such that the absolute values of the leading principal minors $\Delta_1, \Delta_2, \dots, \Delta_n$ of $J(x)$ satisfy

$$|\Delta_1| \geq \epsilon, \frac{|\Delta_2|}{|\Delta_1|} \geq \epsilon, \dots, \frac{|\Delta_n|}{|\Delta_{n-1}|} \geq \epsilon$$

for all $x \in \mathbb{R}^n$, then H is one-to-one from \mathbb{R}^n onto \mathbb{R}^n .

The condition stated on the absolute values of leading principal minors is called the *ratio condition*.

Theorem 4.2: Assume that the controllability matrix of system (1) is nonsingular on \mathbb{R}^n , the set $\{g, [f, g], \dots, (ad^{n-2}f, g)\}$ is involutive on \mathbb{R}^n , and the noncharacteristic matrix satisfies the ratio condition on \mathbb{R}^n . Then there exists a \mathcal{C}^∞ transformation $T = (T_1, T_2, \dots, T_{n+1})$ with the following properties.

- 1) $T(0) = 0$;
- 2) T_1, T_2, \dots, T_n are functions of x_1, x_2, \dots, x_n only and the $n \times n$ Jacobian matrix is nonsingular at each point of \mathbb{R}^n ;
- 3) T_{n+1} is a function of $(x_1, x_2, \dots, x_n, u)$ which can be inverted as a function of u and where $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$;
- 4) T maps the system (1) to the system (2);
- 5) the map (T_1, T_2, \dots, T_n) is one-to-one on \mathbb{R}^n and T is one-to-one on \mathbb{R}^{n+1} .

Proof: By Theorem 4.1, the map F whose Jacobian matrix is the noncharacteristic matrix is one-to-one from \mathbb{R}^n onto \mathbb{R}^n . Thus, we can globally solve for s, t_1, \dots, t_{n-1} as functions of x_1, x_2, \dots, x_n , and by the construction of $T = (T_1, T_2, \dots, T_{n+1})$ in the preceding section, the transformation (T_1, T_2, \dots, T_n) is defined on all of \mathbb{R}^n . Moreover, since the controllability matrix has rank n , the map (T_1, T_2, \dots, T_n) is nonsingular on \mathbb{R}^n as mentioned previously. Properties 1), 3), and 4) follow by the design of T .

We know that the transformation (T_1, T_2, \dots, T_n) has a nonsingular Jacobian matrix with respect to s, t_1, \dots, t_{n-1} and that all entries above the diagonal are zero by comments in Section III. By construction we also have $T_1 = s$, T_2 is a function of s and t_1 only with $\partial T_2 / \partial t_1 = 0$ in \mathbb{R}^n , \dots , T_{n-1} is a function of s, t_1, \dots, t_{n-2} only with $\partial T_{n-1} / \partial t_{n-2} = 0$ on \mathbb{R}^n , and T_n satisfies $\partial T_n / \partial t_{n-1} = 0$ on all of \mathbb{R}^n . This implies that (T_1, T_2, \dots, T_n) is one-to-one on \mathbb{R}^n . The fact that $T = (T_1, T_2, \dots, T_{n+1})$ is one-to-one on \mathbb{R}^{n+1} follows from $(-1)^{n-1} \langle dT_1, (ad^{n-1}f, g) \rangle u + \langle dT_n, f \rangle = T_{n+1}$ [see (5) and (6)].

Example 4.1: Consider the nonlinear system on \mathbb{R}^2 :

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}x_1^2 + e^{x_2} + x_2 - 1 \\ x_1^2 \end{bmatrix} + u \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ = f(x(t)) + u(t)g(x(t)).$$

Computing, we find

$$[f, g] = \begin{bmatrix} -(e^{x_2} + 1) \\ 0 \end{bmatrix}$$

which is linearly independent of g on \mathbb{R}^2 . We first solve $dx_1/ds = -(e^{x_2} + 1)$ with $x_1(0) = 0$ and $dx_2/ds = 0$ with $x_2(0) = 0$ to obtain $x_1 = -2s$ and $x_2 = 0$. Next, we solve $dx_1/dt = 0$ with $x_1(s, 0) = -2s$ and $dx_2/dt = 1$ with $x_2(s, 0) = 0$ to obtain $x_1 = -2s$ and $x_2 = t$.

The function $T_1 = s$ certainly satisfies (6) as a function of s and t . The noncharacteristic matrix is

$$\begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}$$

which fulfills the ratio condition with $\epsilon = 1/2$. Hence our transformation is defined on all of \mathbb{R}^2 , the Jacobian matrix

of (T_1, T_2) is nonsingular everywhere. (T_1, T_2) is one-to-one on \mathbb{R}^2 , and $T = (T_1, T_2, T_3)$ is one-to-one on \mathbb{R}^3 .

Since $T_1 = s$, we have

$$T_2 = \frac{\partial T_1}{\partial s} \frac{\partial s}{\partial x_1} \left(\frac{1}{2}(-2s)^2 + e^{t-1} + t \right) + \frac{\partial T_1}{\partial s} \frac{\partial s}{\partial x_2} (-2s)^2 \\ = -\frac{1}{2}(2s^2 + e^{t-1} + t).$$

The Jacobian matrix of (T_1, T_2) with respect to (s, t) is

$$\begin{bmatrix} 1 & 0 \\ -2s & -\frac{1}{2}(e^t + 1) \end{bmatrix}$$

which satisfies the ratio condition on \mathbb{R}^2 with $\epsilon = 1/2$. Thus, the map is also onto \mathbb{R}^2 .

In Theorem 4.1, the same conclusion holds if the ratio condition on $J(x)$ is replaced by the ratio condition on $AJ(x)$ for some nonsingular constant $n \times n$ matrix A (see [16]).

Corollary 4.1: Suppose that there exists an $n \times n$ nonsingular constant matrix A such that A multiplied on the right by the noncharacteristic matrix satisfies the ratio condition on \mathbb{R}^n . If the controllability matrix of system (1) is nonsingular on \mathbb{R}^n and if the set $\{g, [f, g], \dots, (ad^{n-2}f, g)\}$ is involutive on \mathbb{R}^n , then the conclusions of Theorem 4.2 hold.

Given a nonlinear system (1), it is often possible to determine that there is not a global transformation T with (T_1, T_2, \dots, T_n) mapping \mathbb{R}^n in a one-to-one manner without going through the process of showing that one cannot be constructed. Remembering that T_{n+1} is the control in (2), we write that linear system as

$$\begin{bmatrix} \dot{T}_1 \\ \dot{T}_2 \\ \vdots \\ \dot{T}_n \end{bmatrix} = \begin{bmatrix} T_2 \\ T_3 \\ \vdots \\ 0 \end{bmatrix} + T_{n+1} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = A\tilde{T} + vB \quad (8)$$

where the definitions of \tilde{T} , v , and B are obvious.

Lemma 4.1: Under a transformation T the set of points where f and g are linearly dependent must map by (T_1, T_2, \dots, T_n) to the set of points where $A\tilde{T}$ and B are linearly dependent.

Proof: The set of points where $A\tilde{T}$ and B are linearly dependent is defined by $T_2 = 0, T_3 = 0, \dots, T_n = 0$. If x is a point in \mathbb{R}^n such that $f(x) = cg(x)$ for some constant c , then from (4) we have

$$\langle dT_i, g \rangle = 0, \quad i = 1, 2, \dots, n-1$$

which implies that $\langle dT_i, cg \rangle(x) = 0$ or $\langle dT_i, f \rangle(x) = 0$. Then $\langle dT_i, f \rangle = T_{i+1}$, $i = 1, 2, \dots, n-1$, giving us $T_2 = 0, T_3 = 0, \dots, T_n = 0$ at x . Q.E.D.

The proof of this lemma actually shows the invariance of the linearly dependent sets under the kind of transformations we consider. In [19], [20], and [21] the importance of these sets in controllability is demonstrated. Given a point

in the linearly dependent set of $\dot{x} = f - ug$, feedback can be used to take this point to an equilibrium point.

Lemma 4.1 can be applied to show the impossibility of having a global one-to-one transformation in certain cases.

Example 4.2: We take the nonlinear system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 - x_2^2 \\ 0 \end{bmatrix} + u \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ = f(x(t)) + u(t)g(x(t))$$

on \mathbb{R}^2 . Using the techniques of [23], it can be shown that this is a controllable nonlinear system on \mathbb{R}^2 . Is there a global transformation $T = (T_1, T_2, T_3)$ with a nonsingular Jacobian matrix and with (T_1, T_2) mapping \mathbb{R}^2 in a one-to-one fashion so that

$$\dot{T}_1 = T_2$$

$$\dot{T}_2 = T_3?$$

The set of points where f and g are linearly dependent consists of the two straight lines, $x_2 = 0$ and $x_2 = 1$. In the linear system the vector fields

$$\begin{bmatrix} T_2 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

are linearly dependent if and only if $T_2 = 0$. There is certainly no one-to-one map taking two straight lines to one. Hence, this system is not globally transformable in a one-to-one sense.

It is possible to construct a transformation in a neighborhood of the origin. The solution to

$$\frac{dx}{ds} = [f, g] = \begin{bmatrix} 2x_2 - 1 \\ 0 \end{bmatrix}$$

with $x_1(0) = 0$ and $x_2(0) = 0$ is $x_1 = -s$ and $x_2 = 0$. Solving

$$\frac{dx}{dt} = g = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

with $x_1(s, 0) = -s$ and $x_2(s, 0) = 0$, we have $x_1 = -s$ and $x_2 = t$. The noncharacteristic matrix is

$$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},$$

and this satisfies the ratio condition on \mathbb{R}^2 . Thus, our map $(x_1(s, t), x_2(s, t))$ is a one-to-one transformation of \mathbb{R}^2 onto \mathbb{R}^2 .

The controllability matrix

$$\begin{bmatrix} 0 & 2x_2 - 1 \\ 1 & 0 \end{bmatrix}$$

has rank 2 on $\{(x_1, x_2): x_2 < 1/2\}$. From [5], we know that a transformation exists in a neighborhood of the origin in (x_1, x_2) space. Setting $T_1 = s$, $x_1 = -s$, and $x_2 = t$, we have the transformation $T_1 = -x_1$, $T_2 = -x_2 + x_2^2$, and $T_3 = (1 - 2x_2)u$, which exists on the set $\{(x_1, x_2): x_2 < 1/2\}$. Moreover, the Jacobian matrix of (T_1, T_2) is nonsingular on this set and (T_1, T_2) is one-to-one there.

This example suggests the following corollary to Theorem 4.2 (and Corollary 4.1), the proof of which follows that of the theorem. Let U be an open set in \mathbb{R}^n containing the origin and f and g be \mathcal{C}^∞ on \mathbb{R}^n .

Corollary 4.2: Assume that the controllability matrix of system (1) is nonsingular on U , the set $\{g, [f, g], \dots, (ad^{n-2}f, g)\}$ is involutive on U , and the noncharacteristic matrix (or the noncharacteristic matrix with a premultiplication by an $n \times n$ nonsingular constant matrix A) satisfies the ratio condition on \mathbb{R}^n . Then there exists a \mathcal{C}^∞ transformation $T = (T_1, T_2, \dots, T_{n-1})$ with the properties 1)-5) of Theorem 4.2 holding where \mathbb{R}^n is replaced by U .

V. OTHER TRANSFORMATION RESULTS

Suppose that f and g in (1) are \mathcal{C}^∞ on \mathbb{R}^n . The proof of Theorem 4.1 in [16] applies to any n -dimensional open (possibly unbounded) rectangle R in (x_1, x_2, \dots, x_n) space with sides parallel to one of $x_1 = 0, x_2 = 0, \dots, x_n = 0$ and containing the origin [in this case we have H is one-to-one onto the image $H(R)$]. If U is an open subset of \mathbb{R}^n containing the origin and $U \subset F^{-1}(R)$ (the inverse image of R under the map F whose Jacobian is the noncharacteristic matrix) and if the noncharacteristic matrix, with a possible nonsingular premultiplication, satisfies the condition on R , then a result similar to Corollary 4.2 can be proved.

We are interested in assumptions other than the ratio condition which can be applied to the noncharacteristic matrix. The following result is found in [16].

Theorem 5.1: Suppose that there is a map $H: \mathbb{R}^n \rightarrow \mathbb{R}^n$ which is defined on an open convex subset Ω of \mathbb{R}^n .

a) If H is differentiable with Jacobian matrix $J(x)$, and if there exists a constant $n \times n$ nonsingular matrix A so that $AJ(x)$ is positive-definite for all $x \in \Omega$, then H is one-to-one from Ω onto $H(\Omega)$.

b) If H is continuously differentiable with Jacobian matrix $J(x)$, if Ω is bounded, and if there exists a constant $n \times n$ nonsingular matrix A so that $\det AJ(x) > 0$ for all $x \in \Omega$ and $AJ(x) + (AJ(x))^*$ has nonnegative principal minors for all $x \in \Omega$, then H is one-to-one from Ω onto $H(\Omega)$.

In our application of this theorem we assume that Ω contains an open neighborhood of the origin in (s, t_1, \dots, t_{n-1}) space. The proof of our next result is obvious from previous considerations.

Theorem 5.2: Let U be an open set in \mathbb{R}^n containing the origin and with $U \subset f^{-1}(\Omega)$. Suppose that the controllability matrix of system (1) is nonsingular on U , the set $\{g, [f, g], \dots, (ad^{n-2}f, g)\}$ is involutive on U , and the noncharacteristic matrix satisfies a) or b) in Theorem 5.1. Then there exists a transformation satisfying the following conclusions.

1) $T(0) = 0$;

2) T_1, T_2, \dots, T_n are functions of x_1, x_2, \dots, x_n only and the $n \times n$ Jacobian matrix is nonsingular on U .

3) T_n is a function of x_1, x_2, \dots, x_n, u which can be inverted as a function of u and where $(x_1, x_2, \dots, x_n) \in U$;

4) T maps the system (1) to the system (2);

5) the map (T_1, T_2, \dots, T_n) is one-to-one on U and T is one-to-one on $U \times R$.

This result, using condition a),

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},$$

$\Omega = R^2$, and $U = \{(x_1, x_2): x_2 < 1/2\}$ can be applied to Example 4.2. If we let Ω be any bounded square about the origin in (s, t) space, U be the same square in (x_1, x_2) space intersected with $\{(x_1, x_2): x_2 < 1/2\}$, and

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},$$

then condition b) applies.

At a CBMS Conference in 1978 at the University of California, Davis, Brockett discussed problems concerning the mapping of nonlinear systems and a construction like that given here. We remark that our Assumption ii) in Section II is weaker than a similar assumption in [2]. In that paper the hypothesis is that

$$[(ad^k f, g), (ad^j f, g)] = \sum_{i=0}^{d-1} c_i(ad^i f, g),$$

$$0 \leq k, j \leq n-2$$

where $d = \max(j, k)$ and the c_i are functions. For example, the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \sin x_2 \\ 0 \end{bmatrix} + u \begin{bmatrix} \cos x_1 \\ 1 \end{bmatrix}$$

does not satisfy this condition, but is transformable in our theory since ii) is trivially satisfied.

We remark that Theorem 4.2 (and similar results) requires the construction of the noncharacteristic matrix. Explicit constructions are possible for block-triangular systems [3], [4]. Ford is working on building transformations using the symbolic computation of the M.I.T. MACSYMA Program. This appears to be theoretically possible for those systems that satisfy Brockett's [2] conditions and bilinear systems satisfying the assumptions of Su [5]. Also, other systems can be handled, and numerical techniques are being investigated.

REFERENCES

- [1] A. J. Krener, "On the equivalence of control systems and the linearization of nonlinear systems," *SIAM J. Contr.*, vol. 11, pp. 670-676, 1973.
- [2] R. W. Brockett, "Feedback invariants for nonlinear systems," IFAC Congr., Helsinki, Finland, 1978.
- [3] G. Meyer and L. Cicolani, "A formal structure for advanced flight control systems," NASA TN D-7940, 1975.
- [4] G. Meyer and L. Cicolani, "Application of nonlinear system inverses to automatic flight control design—System concepts and flight evaluations," in *Theory and Applications of Optimal Control in Aerospace Systems*, P. Kanti, Ed., AGARDograph, 1980.
- [5] R. Su, "On the linear equivalents of nonlinear systems," *Syst. Contr. Lett.*, vol. 2, pp. 48-52, 1982.
- [6] J. Hadamard, "Sur les transformations planes," *C. R. Acad. Sci.*, vol. 142, pp. 71-84, 1906.
- [7] J. Hadamard, "Sur les transformations ponctuelles," *Bull. Soc. Math. France*, vol. 34, pp. 77-84, 1906.
- [8] J. Hadamard, *Sur Les Correspondences*, Oeuvres, pp. 383-384.
- [9] R. S. Palais, "Natural operations on differential forms," *Trans. Amer. Math. Soc.*, vol. 92, pp. 125-141, 1959.
- [10] M. S. Berger and M. S. Berger, *Perspectives in Nonlinearity*, New York: Benjamin, 1970.
- [11] F. F. Wu and C. A. Desoer, "Global inverse function theorem," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 199-201, 1972.
- [12] E. S. Kuh and I. Hajj, "Nonlinear circuit theory: Resistive networks," *Proc. IEEE*, vol. 59, pp. 340-355, 1971.
- [13] J. M. Fitts, "On the observability of nonlinear systems with applications to nonlinear regression analysis," presented at the Symp. Nonlinear Estimate Theory and Its Appl., San Diego, CA, 1970.
- [14] E. W. Griffith and K. S. P. Kumar, "On the observability of nonlinear systems, I," *J. Math. Anal. Appl.*, vol. 35, pp. 135-147, 1971.
- [15] T. Fujisawa and E. S. Kuh, "Some results on existence in uniqueness of solutions of nonlinear networks," *IEEE Trans. Circuit Theory*, vol. CT-18, pp. 501-506, 1971.
- [16] S. R. Kou, D. L. Elliot, and T. J. Tarn, "Observability of nonlinear systems," *Inform. Contr.*, vol. 22, pp. 89-99, 1973.
- [17] I. W. Sandberg, "Global implicit function theorems," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 145-149, 1981.
- [18] J. D. Miller, "Global inverse function theorem," submitted for publication.
- [19] L. R. Hunt, "N-dimensional controllability with $n-1$ controls," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 113-116, 1982.
- [20] L. R. Hunt, "Sufficient conditions for controllability," *IEEE Trans. Circuits Syst.*, vol. CAS-29, pp. 285-288, 1982.
- [21] L. R. Hunt, "Global controllability of nonlinear systems in two dimensions," *Math. Syst. Theory*, vol. 13, pp. 361-376, 1980.
- [22] R. Hermann and A. J. Krener, "Nonlinear controllability and observability," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 728-740, 1977.
- [23] R. W. Brockett, "Nonlinear systems and differential geometry," *Proc. IEEE*, vol. 64, pp. 61-72, 1976.
- [24] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, New York: Academic, 1975.
- [25] L. H. Loomis and S. Sternberg, *Advanced Calculus*, Reading, MA: Addison-Wesley, 1968.
- [26] L. Auslander and R. E. MacKenzie, *Introduction to Differentiable Manifolds*, New York: Dover, 1977.
- [27] A. Isidori, A. J. Krener, C. Gori-Giorgi, and S. Monaco, "Nonlinear decoupling via feedback: A differential geometric approach," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 331-345, 1981.
- [28] F. John, *Partial Differential Equations*, New York: Springer-Verlag, 1971.



L. R. Hunt (A'79-M'81) was born in Shreveport, LA, on December 5, 1942. He received the B.S. degree in mathematics from Baylor University, Waco, TX, in 1964 and the Ph.D. degree in mathematics from Rice University, Houston, TX, in 1970.

He has been a faculty member in the Department of Mathematics, Texas Tech University, Lubbock, since 1969, and he currently holds the rank of Professor. He is a Senior Investigator for the Joint Services Electronics Program, Department of Electrical Engineering, Texas Tech University. From 1980 to 1982 he was on leave under an Intergovernmental Personnel Agreement at NASA Ames Research Center, Moffett Field, CA. His research interests are nonlinear systems and control, several complex variables, and partial differential equations.

Dr. Hunt is a member of the American Mathematical Society and the Society for Industrial and Applied Mathematics.

Renjeng Su was born in China in 1950. He received the B. S. degree in chemical engineering from Chen-Kung University, Taiwan, in 1972 and the M. S. and D. Sc. degrees, both in systems science and mathema-



tica, from Washington University, St. Louis, MO, in 1977 and in 1980, respectively.

From 1980 to 1982 he worked at NASA Ames Research Center, Moffett Field, CA, as a Research Associate of the National Research Council, and in August 1982 he joined the Department of Electrical Engineering, Texas Tech University, Lubbock, as an Assistant Professor. His main research interests are in nonlinear systems and control.



George Meyer (M61--M78) was born in Panchievo, Yugoslavia, on October 25, 1934. He received the B.S., M.S., and Ph.D. degrees, all from the University of California, Berkeley, in 1957, 1960, and 1965, respectively.

He has been employed by the Boeing Aeroplane Company, Seattle, WA, and Ampex Corporation, Redwood City, CA. Since 1965 he has been with the NASA Ames Research Center, Moffett Field, CA. His primary area of research is automatic control.

n-Dimensional Controllability with (n-1) Controls

LOUIS R. HUNT

Abstract—Let M be a connected real-analytic n -dimensional manifold, f, g_1, \dots, g_{n-1} be complete real-analytic vector fields on M which are linearly independent at some point of M , and u_1, \dots, u_{n-1} be real-valued controls. Consider the controllability of the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M.$$

Necessary and sufficient conditions are given so that this system is controllable on any simply connected domain D contained in M on which g_1, \dots, g_{n-1} are linearly independent. These conditions depend on the computation of Lie brackets at those points where f, g_1, \dots, g_{n-1} are linearly dependent.

1. INTRODUCTION

Consider the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{n-1} u_i(t) g_i(x(t)), \quad x(0) = x_0 \in M \quad (1)$$

where M is a connected real-analytic n -dimensional paracompact manifold, and f, g_1, \dots, g_{n-1} are complete real-analytic vector fields on M , which are linearly independent at some point in M . Let D be a simply connected domain in M on which g_1, \dots, g_{n-1} are linearly independent. We are interested in establishing necessary and sufficient conditions that the system (1) be controllable on D . These conditions involve the computation of the Lie brackets of f, g_1, \dots, g_{n-1} at those points where $f, g_1, g_2, \dots, g_{n-1}$ are linearly dependent and where the vector space dimension of the Lie algebra L_A generated by g_1, g_2, \dots, g_{n-1} and successive Lie brackets is $(n-1)$.

In the special case $n=2$, necessary and sufficient conditions for controllability are given in [10] for the system

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)), \quad x(0) = x_0 \in M. \quad (2)$$

Let $(ad_g, f) = [g, f]$, the Lie bracket of f and g , $(ad^2_g, f) = [g, [g, f]]$, etc. Then the system (2) is controllable on D from any point $x_0 \in D$, where D is a simply connected domain in M on which g is nonzero if and only if

Manuscript received June 27, 1980; revised May 12, 1981. Paper recommended by M. Vidyasagar, Past Chairman of the Stability, Nonlinear, and Distributed Systems Committee. This work was supported by the National Science Foundation under Grant MCS 76-02257-A01 and by the Joint Services Electronics Program under ONR Contract N00014-76-C-1134.

The author is with the Department of Mathematics, Texas Tech University, Lubbock, TX 79409.

the following condition holds. Every integral curve of g which disconnects D contains a point p where f and g are linearly dependent and the first integer j such that (ad^j_g, f) and g are linearly independent at p is odd. Also assumed here is that the vector space dimension of the Lie algebra generated by f, g and successive Lie brackets is 2 at x_0 in order that we know an open subset of D is reachable from x_0 .

Let L_A be the Lie algebra generated by f, g_1, \dots, g_{n-1} in (1) and successive Lie brackets, and suppose that the vector space dimension of L_A at x_0 is n . Then an open subset of M is reachable from x_0 by results in [9] and [11], and we denote by U the largest open subset in M which is reachable from x_0 called the region of reachability from x_0 . In [1], [6], [7], and [8] this set U (which is shown to be connected in [7]) is characterized in the following way. The boundary of $U, \partial U$, consists of the $(n-1)$ -dimensional integral manifolds of g_1, \dots, g_{n-1} (and hence of the Lie algebra L_A) that intersect it. Moreover, the integral curve of f which starts at a point in ∂U must remain in \bar{U} , the closure of U in M .

Hence, if there are no integral manifolds of L_A in D or if every integral manifold N of D separates D into two open disjoint sets O_1 and O_2 with $O_1 \cup O_2 \cup N = D$ and contains points where the vector field f is "in the directions of O_1 ," and other points where the vector field f is "in the direction of O_2 ," then D is controllable from any point $x_0 \in D$ with the vector space dimension of L_A at x_0 equal n . The problem is to determine this turning of the vector field f through the integral manifold N . This must occur at a point p of N where f and g_1, \dots, g_{n-1} are linearly dependent, i.e., a point p where f is in the tangent space to N at p . Thus, we have a local question and we can apply the beautiful work of Hermes [3] to study the turning of f through N . The answers are obtained in terms of Lie brackets at p . We are using unbounded controls in our theory, and locally this has the effect of applying "impulsive controls" in [3].

Section II of this paper contains definitions and known results. In Section III we give examples and prove our main theorem.

II. DEFINITIONS

We are concerned with the controllability of the system given in (1).

If h_1 and h_2 are C^∞ vector fields on M , we define the Lie bracket of h_1 and h_2 by

$$[h_1, h_2] = \frac{\partial h_2}{\partial x} h_1 - \frac{\partial h_1}{\partial x} h_2$$

where $\partial h_1 / \partial x$ and $\partial h_2 / \partial x$ are Jacobian matrices. For our purpose, we could have taken $[h_1, h_2]$ to be the negative of the above expression as well. Of course it is possible to define higher order Lie brackets $[h_1, [h_1, h_2]], [h_2, [h_1, h_2]], \dots$. We set $(adh_1, h_2) = [h_1, h_2]$ and inductively $(ad^{k+1}h_1, h_2) = [h_1, (ad^k h_1, h_2)]$. Let L_A and L'_A be the Lie algebras generated by f, g_1, \dots, g_{n-1} and successive Lie brackets and g_1, \dots, g_{n-1} and successive Lie brackets, respectively, where f, g_1, \dots, g_{n-1} are defined in (1). For $x \in M$, we set $L_A(x) = \{h(x) : h \in L_A\}$ and $L'_A(x) = \{h(x) : h \in L'_A\}$. The set of vector fields g_1, \dots, g_{n-1} is called involutive at x if there exist functions γ_{ijk} such that

$$[g_i, g_j](x) = \sum_{k=1}^{n-1} \gamma_{ijk}(x) g_k(x) \quad \text{for all } i, j, 1 \leq i, j \leq n-1, i \neq j.$$

If this holds for all $x \in M$, then g_1, \dots, g_{n-1} are involutive on M .

By $T(M)$ we denote the tangent bundle to M with fiber $T_x(M)$ for $x \in M$, the tangent space to M at x . If X is a C^∞ vector field on M , then α is an integral curve of X if α is a C^∞ mapping from an interval $I \subset \mathbb{R}$ into M such that $d\alpha(t)/dt = X(\alpha(t))$ for all $t \in I$. For S a subset of $T(M)$, an integral curve of S is a mapping α from a real interval $[t_0, t_1]$ into M so that there exist $t_0 < t_1 < t_2 < \dots < t_k = t_1$ and vector fields X_1, \dots, X_k in S with the restriction of α to $[t_{i-1}, t_i]$ being an integral curve of X_i for each $i = 1, 2, \dots, k$. A connected submanifold N of M is an integral manifold of g_1, \dots, g_{n-1} if $T_x(N)$ is the space spanned by g_1, \dots, g_{n-1} at y for each $y \in N$.

The subset S of M we consider is the one given by the vector fields in our system (1). A point $x \in M$ is reachable from $x_0 \in M$ if there is an

integral curve α of S and some time $T \geq 0$ in the interval for α such that $\alpha(0) = x_0$ and $\alpha(T) = x$. A subset A of M is *reachable* from x_0 if every $x \in A$ is reachable from x_0 . In the case $A = M$, the system is *controllable* from x_0 . If the system is controllable from every x_0 in M , then it is *controllable*.

We apply the above definitions to our simply connected domain D in M on which g_1, \dots, g_{n-1} are linearly independent. In fact, since our main results are stated for D , we may as well assume that $M = D$. Only in the examples following Theorem 3.1 in Section III will we consider cases where D is properly contained in M .

Letting y be a point in D , we need to examine the possibility of having an integral manifold N of g_1, \dots, g_{n-1} through y . The real-analytic version of the famous Frobenius theorem states if g_1, \dots, g_{n-1} are involutive at y (or equivalently the vector space dimension of $L'_y(y)$ is $(n-1)$), then there is a unique real-analytic $(n-1)$ -dimensional integral manifold N of g_1, \dots, g_{n-1} through y . If there is a point $y' \in \partial N \cap D$, then the vector fields g_1, \dots, g_{n-1} are involutive at y' by continuity and $y' \in N \cap D$. Thus, an $(n-1)$ -dimensional integral manifold of g_1, \dots, g_{n-1} in D can have no boundary points in D . Of course, if the vector space dimension of $L'_y(y)$ is n , then there is no integral manifold of g_1, \dots, g_{n-1} of dimension $(n-1)$ through y , and an open neighborhood O of y in D is reachable from any point in O by Chow's theorem (see [11]).

We emphasize that we are not assuming that the vector space dimension of L'_y is a constant on D . It can vary as we move through D , but must always be either $(n-1)$ or n . Points where this dimension is $(n-1)$ are interior points of integral manifolds of g_1, \dots, g_{n-1} . We say that an integral manifold N of g_1, \dots, g_{n-1} *disconnects* D if there are disjoint, nonempty open sets O_1 and $O_2 = \text{complement in } D \text{ of } (O_1 \cup N)$ with $O_1 \cup O_2 \cup N = D$ (i.e., N separates D into disjoint, nonempty open sets O_1 and O_2). In view of the preceding paragraph and the fact that D is simply connected, the following assumption is reasonable. For the remainder of this paper, we assume that every integral manifold of g_1, \dots, g_{n-1} in D disconnects D .

Let V be an open subset of D and take $x \in \partial V$. Then f points in the direction of \bar{V} (or toward \bar{V}) at x if there is an open neighborhood W of x in D , such that the integral curve of f starting at x and intersected with W is contained in \bar{V} . If this is true for all $x \in \partial V$, then f points in the direction of \bar{V} on ∂V . We define f pointing in the direction of V by replacing \bar{V} everywhere with V .

Our first two results are proved in [6].

Theorem 2.1: Assume that the vector space dimension of L'_x at $x_0 \in D$ is n . Let U be the largest open subset of D which is reachable from x_0 . Then ∂U consists of integral manifolds of g_1, \dots, g_{n-1} and f points in the direction of \bar{U} on ∂U .

If there is an integral manifold N of g_1, \dots, g_{n-1} disconnecting D into O_1 and O_2 with f pointing toward \bar{O}_1 (or \bar{O}_2) on N , then the system is certainly not controllable on D by results in [6].

Let Λ^d denote Hausdorff measure (see [2]) in dimension d on D . Suppose I is the set of points on which the Lie algebra L'_x has dimension $(n-1)$. Then any integral manifold N of g_1, \dots, g_{n-1} is contained in I , and for such a manifold we must have $\Lambda^{n-1}(L) > 0$.

Theorem 2.2: If $\Lambda^{n-1}(L) = 0$ then our system (1) is controllable from any $x_0 \in D$.

Next we turn to the results of Hermes [3] concerning local controllability along a reference solution. Let p be a point in D where $f(p) = 0$, f being as in our system (1). That is, p is a critical point or rest solution when all controls are zero. Suppose that the vector space dimension of $L'_p(p)$ is $(n-1)$. We note that g_1, \dots, g_{n-1} do not have to be linearly independent in [3], but we assume they are for now and discuss how this assumption can be removed for our entire theory later in this paper. We also remark that the results in [3] are stronger than indicated here.

Let $\langle \cdot, \cdot \rangle$ denote the inner product of tangent vectors induced by a Riemannian metric on D . By h we denote a vector field on D so that g_1, \dots, g_{n-1}, h are linearly independent at p . Let l be the unique vector such that

$$\langle l(p), h(p) \rangle = 1, \quad \langle l(p), g_i(p) \rangle = 0, \quad i = 1, \dots, n-1.$$

Set $v = (v_1, \dots, v_{n-1})$, r , a nonnegative integer

$$v^r = \sum_{i=1}^{n-1} v_i \cdot v^i = v_1^r \cdot \dots \cdot v_{n-1}^r$$

$$a(v) = (l(p), (ad^{r-1}g_{n-1}, (ad^{r-2}g_{n-2}, (\dots, (ad^{r-1}g_1, f), \dots)(p))),$$

$$\phi_r(s) = \sum_{i=1}^{n-1} \frac{1}{i!} (-s_1)^i \dots (-s_{n-1})^{i_{n-1}} a(v)$$

where (s_1, \dots, s_{n-1}) is any point in a neighborhood of the origin in \mathbb{R}^{n-1} .

The following theorem can be deduced from the results in [3].

Theorem 2.3: Consider the system (1) with $f(p) = 0$ and with the vector space dimension of $L'_p(p)$ equal $(n-1)$, and let N be the integral manifold of g_1, \dots, g_{n-1} (and hence L'_p) through p . Assume that there exists an integer $r \geq 1$ such that $\phi_r(s) \neq 0$ for s in some neighborhood of the origin in \mathbb{R}^{n-1} , and let r^* be the smallest integer with this property.

1) If O is any open neighborhood of p in D , let N disconnect O into two disjoint open sets O_1 and O_2 . If $\phi_{r^*}(s)$ changes sign in every neighborhood of the origin in \mathbb{R}^{n-1} , then there exist $y_1, y_2 \in N \cap O$, such that f points toward O_1 at y_1 and toward O_2 at y_2 .

2) If O is any open neighborhood of p in D , let N, O_1 , and O_2 be as above. Then, if $\phi_{r^*}(s)$ is semidefinite, $\sum_{i=1}^{r^*} \phi_i(s)$ changing sign in every neighborhood of the origin in \mathbb{R}^{n-1} is a necessary and sufficient condition that there exist $y_1, y_2 \in N \cap O$, such that f points toward O_1 at y_1 and toward O_2 at y_2 .

The task is to merge Theorems 2.1, 2.2, and 2.3 to give necessary and sufficient conditions for our system to be controllable on D .

III. RESULTS AND EXAMPLES

We first prove the main theorem and then give a number of illustrative examples. Recall that we are assuming g_1, \dots, g_{n-1} are linearly independent on \bar{D} , all $(n-1)$ -dimensional integral manifolds of g_1, \dots, g_{n-1} disconnect D , and f, g_1, \dots, g_{n-1} are linearly independent at some point of D .

If the vector space dimension of L'_x is n at every point of D , then the system is controllable on D by Theorem 2.2 (see [4], [10]). If the vector space dimension of L'_x is $(n-1)$ at some point x of D , then $L'_x(x) = L'_x(x)$ and by Nagano's theorem there is a $(n-1)$ -dimensional integral manifold of f, g_1, \dots, g_{n-1} through this point which disconnects D . It is impossible to move from one side of this manifold in D to the other, so this system is not controllable by the comment after Theorem 2.1.

Thus, we assume that the dimension of L'_x is n at every point of D and the dimension of L'_x is $(n-1)$ at some point in D (i.e., we have at least one integral manifold of L'_x (of g_1, \dots, g_{n-1}) in D). Suppose that f, g_1, \dots, g_{n-1} are linearly independent at every point of N . Then we have an integral manifold disconnecting D into two disjoint open sets O_1 and O_2 such that f points in the direction of one of the sets, say O_1 , on N , and the system is not controllable by the comment after Theorem 2.1.

Hence, we additionally assume that every integral manifold N of g_1, \dots, g_{n-1} in D contains a point p where f, g_1, \dots, g_{n-1} are linearly dependent. Suppose that c_1, \dots, c_{n-1} are constants such that

$$f(p) = c_1 g_1(p) + \dots + c_{n-1} g_{n-1}(p).$$

If we replace u_1 by $u_1 - c_1$, u_2 by $u_2 - c_2, \dots, u_{n-1}$ by $u_{n-1} - c_{n-1}$ in system (1), then we have a new "equivalent" system

$$\begin{aligned} \dot{x}(t) = & (f - (c_1 g_1 + c_2 g_2 + \dots + c_{n-1} g_{n-1}))(x(t)) \\ & + \sum_{i=1}^{n-1} u_i(t) g_i(x(t)). \end{aligned} \quad (3)$$

This new system has the property that p is a critical point for the drift term (i.e., when u_1, \dots, u_{n-1} are all 0). Moreover, Theorem 2.3 is applicable to system (3). Since

$$\begin{aligned} [f - (c_1 g_1 + c_2 g_2 + \dots + c_{n-1} g_{n-1}), g_i] \\ = [f, g_i] - c_1 [g_1, g_i] - \dots - c_{n-1} [g_{n-1}, g_i] \end{aligned}$$

and $[f, g_i]$ is in the tangent bundle to N , g_1, \dots, g_{n-1} being involutive, the Lie brackets $[f, g_i]$ and $[f, (c_1 g_1 + c_2 g_2 + \dots + c_{n-1} g_{n-1}), g_i]$ differ by vector fields in the tangent bundle to N . This is also true for all successive Lie brackets of interest to us. We can define the functions ϕ_i and ϕ_{i-} associated with Theorem 2.3 for our system (1) where f is tangent to N at p . Since we are interested in the direction of f on N near p , the information given to us by Theorem 2.3 applied to system (3) is the same as that given by direct calculations from system (1). In fact, the numbers $a(p)$ given by system (1) and system (3) agree for all p because the inner product of $f(p)$ with any tangent vector to N at p is 0.

Thus, we have proved the following theorem in which ϕ_i and ϕ_{i-} are as in Theorem 2.3.

Theorem 3.1: The system (1) is controllable if any one of the following conditions is satisfied.

- 1) The vector space dimension of L_A is n at every point of D .
- 2) Every integral manifold N of g_1, \dots, g_{n-1} in D contains a point p where f, g_1, \dots, g_{n-1} are linearly dependent and $\phi_{i-}(s)$ changes sign in every neighborhood of the origin in \mathbb{R}^{n-1} .
- 3) Every integral manifold N of g_1, \dots, g_{n-1} in D contains a point p where f, g_1, \dots, g_{n-1} are linearly dependent, $\phi_{i-}(s)$ is semidefinite on some neighborhood of the origin in \mathbb{R}^{n-1} , and $\sum_{i=1}^{n-1} \phi_i(s)$ changes sign in every neighborhood of the origin in \mathbb{R}^{n-1} .

The system (1) is not controllable if any one of the following conditions is satisfied.

- 4) The vector space dimension of L_A is $(n-1)$ at some point of D .
- 5) There is an integral manifold N of g_1, \dots, g_{n-1} on which f, g_1, \dots, g_{n-1} are linearly independent.
- 6) There is an integral manifold N of g_1, \dots, g_{n-1} so that for every point $p \in N$ at which f, g_1, \dots, g_{n-1} are linearly dependent there is an open neighborhood of the origin in \mathbb{R}^{n-1} on which $\phi_{i-}(s)$ is semidefinite and $\sum_{i=1}^{n-1} \phi_i(s)$ does not change sign.

Throughout this paper we have assumed that g_1, \dots, g_{n-1} are linearly independent on D . Since Theorem 2.1 (see [7]) and Theorem 2.3 are valid if we replace this assumption by the hypothesis that the vector space dimension of L_A on D is $(n-1)$, Theorem 3.1 again holds.

We present a series of examples which demonstrate the application of Theorem 3.1 to a variety of systems. Examples for dimension 2 are given in [5].

Example 3.1: We take $M = D = \mathbb{R}^3$ and

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = f(x(t)) + u_1(t) \begin{bmatrix} x_2 \\ x_3 \\ x_1 \end{bmatrix} + u_2(t) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\ = f(x(t)) + u_1(t) g_1(x(t)) + u_2(t) g_2(x(t))$$

where f is any real-analytic vector field on \mathbb{R}^3 . Since g_2 ,

$$[g_1, g_2] = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

and

$$[g_1, [g_1, g_2]] = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

are linearly independent on \mathbb{R}^3 , the vector space dimension of L_A on \mathbb{R}^3 is 3, and part 1) of Theorem 3.1 implies that this system is controllable on \mathbb{R}^3 . Note that part 1) of Theorem 3.1 certainly does not require that g_1 and g_2 be linearly independent on D .

Example 3.2: Let $M = D = \mathbb{R}^3$ and

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x^2 + x_2^2 + x_3^2 \\ 1 \\ 1 \end{bmatrix} + u_1(t) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + u_2(t) \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \\ = f(x(t)) + u_1(t) g_1(x(t)) + u_2(t) g_2(x(t)).$$

The vector fields f, g_1 , and g_2 are linearly dependent if and only if

$x_1^2 - x_2^2 + x_3^2 = 0$, i.e., at the origin. The integral manifolds of g_1 and g_2 are the hyperplanes $x_1 = \text{constant}$. Thus, there are integral manifolds of g_1 and g_2 which do not contain a point where f, g_1 , and g_2 are dependent. By part 5) of Theorem 3.1, this system is not controllable on \mathbb{R}^3 .

Example 3.3: We take $M = D = \mathbb{R}^3$ and

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x_2^2 + x_3^2 \\ 1 \\ 1 \end{bmatrix} + u_1(t) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + u_2(t) \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ = f(x(t)) + u_1(t) g_1(x(t)) + u_2(t) g_2(x(t)).$$

The vector fields f, g_1 , and g_2 are linearly dependent when $x_2^2 + x_3^2 = 0$, i.e., on the x_1 -axis. The integral manifolds of g_1 and g_2 are the hyperplanes $x_1 = \text{constant}$, and each of these intersects the x_1 -axis when $x_2 = x_3 = 0$. Computing Lie brackets we find

$$[f, g_1] = \begin{bmatrix} 2x_3 \\ 0 \\ 0 \end{bmatrix}, \quad [f, g_2] = \begin{bmatrix} 2x_2 \\ 0 \\ 0 \end{bmatrix}, \\ [[f, g_1], g_1] = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \quad [[f, g_1], g_2] = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \\ [[f, g_2], g_1] = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad [[f, g_2], g_2] = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}.$$

Note that both $[f, g_1]$ and $[f, g_2]$ are $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ when $x_2 = x_3 = 0$. Thus, with

$$l(p) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad r^* = 2 \quad \text{in our formulas for } \phi_i$$

$$\phi_{i-}(s) = \frac{1}{2}(-s_1)^2(-s_2)^0 2 + \frac{1}{2}(-s_1)^0(-s_2)^2 2$$

which is positive definite in a neighborhood of (0,0) in \mathbb{R}^2 . By part 6) of Theorem 3.1 this system is not controllable, since

$$\sum_{i=1}^n \phi_i(s) = \phi_{i-}(s).$$

Example 3.4: Let $M = D = \mathbb{R}^3$ and

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + u_1(t) \begin{bmatrix} x_2 \\ x_3 \\ x_1 \end{bmatrix} + u_2(t) \begin{bmatrix} x_3 \\ x_1 \\ x_2 \end{bmatrix} \\ = f(x(t)) + u_1(t) g_1(x(t)) + u_2(t) g_2(x(t)).$$

The vector fields f, g_1 , and g_2 are linearly dependent when $x_1^2 + x_2^2 + x_3^2 - 3x_1x_2x_3 = 0$, and g_1 and g_2 are linearly dependent when $x_1 = x_2 = x_3$. There is a point, e.g., $x_1 = 0, x_2 = -x_3 \neq 0$, where f, g_1 , and g_2 are dependent but g_1 and g_2 are independent. Let D be any simply connected domain in \mathbb{R}^3 containing this point and not intersecting the line $x_1 = x_2 = x_3$. Since

$$[g_1, g_2] = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad [f, g_1] = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$[f, g_2] = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

the dimension of L_A at any point of D is determined by the span of f, g_1 , and g_2 at this point. Hence, there is a point in D where the dimension of L_A is 2, and our system is not controllable by part 4) of Theorem 3.1.

Example 3.5: We take $M = D = \mathbb{R}^3$ and

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x_2^2 - x_3^2 \\ 1 \\ 1 \end{bmatrix} + u_1(t) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + u_2(t) \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$= f(x(t)) + u_1(t)g_1(x(t)) + u_2(t)g_2(x(t)).$$

These three vector fields are linearly dependent if and only if $x_2^2 - x_3^2 = 0$ or $x_2 = x_3$. Computing Lie brackets we find that

$$[f, g_1] = \begin{bmatrix} -2x_1 \\ 0 \\ 0 \end{bmatrix} \text{ and } [f, g_2] = \begin{bmatrix} 2x_1 \\ 0 \\ 0 \end{bmatrix}.$$

Now the integral manifolds of g_1 and g_2 are given by $x_1 = \text{constant}$ and each contains the lines $x_2 = x_3$ when x_1 is held fixed. If we let

$$l(p) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, x_2 = 1 \text{ and } x_3 = -1, \text{ and } r^* = 1 \text{ we have}$$

$$\phi_{r^*}(s) = (-s_1)^1 (s_2)^0 (2) + (-s_1)^0 (s_2)^1 (2)$$

and Theorem 3.1, part 2) gives us global controllability.

Example 3.6: In our preceding example, first-order Lie brackets were applied to imply controllability. Here we use third-order brackets.

Take $M = D = \mathbb{R}^3$ and

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x_2^2 - x_3^2 \\ 1 \\ 1 \end{bmatrix} + u_1(t) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + u_2(t) \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$= f(x(t)) + u_1(t)g_1(x(t)) + u_2(t)g_2(x(t)).$$

We know that f, g_1 , and g_2 are linearly dependent only on the x_1 axis and the plane $x_2 = -x_3$. Every integral manifold of g_1 and g_2 intersects the x_1 axis. Computing when $x_2 = x_3 = 0$ we find that the only Lie brackets of the first three orders that are nonzero are

$$[[[f, g_1], g_1], g_1] = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} \text{ and } [[[f, g_2], g_2], g_2] = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}.$$

Letting $l(p) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $r^* = 3$ we have

$$\phi_{r^*}(s) = \frac{1}{6}(-s_1)^3(-s_2)^0 6 + \frac{1}{6}(-s_1)^0(-s_2)^3 6$$

which satisfies condition 2) in Theorem 3.1 and implies controllability.

If we choose to work at those points where $x_2 = -x_3 \neq 0$, instead of along the x_1 -axis, first-order Lie brackets will suffice.

Example 3.7: Let $M = \mathbb{R}^3 - (x_1\text{-axis})$ and

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -x_2 \\ x_1 \\ 0 \end{bmatrix} + u_1(t) \begin{bmatrix} x_1 \\ -4x_2 \\ 0 \end{bmatrix} + u_2(t) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$= f(x(t)) + u_1(t)g_1(x(t)) + u_2(t)g_2(x(t)).$$

Now g_1 and g_2 are linearly independent on M and f, g_1 , and g_2 are dependent if and only if $x_1 = \pm 2x_2$.

Let

$$D_1 = \{(x_1, x_2, x_3) \in M : x_1, x_2 > 0\}$$

$$D_2 = \{(x_1, x_2, x_3) \in M : x_2 > 0 \text{ and } x_1 < 0\}$$

$$D_3 = \{(x_1, x_2, x_3) \in M : x_1, x_2 < 0\}$$

$$D_4 = \{(x_1, x_2, x_3) \in M : x_1 > 0 \text{ and } x_2 < 0\}.$$

Every integral manifold of g_1 and g_2 in D_1 intersects the line $x_1 = 2x_2$, and D_1 is foliated by such integral manifolds since

$$[g_1, g_2] = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We find that

$$[f, g_1] = -5 \begin{bmatrix} x_2 \\ x_1 \\ 0 \end{bmatrix} \text{ and } [f, g_2] = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Letting

$$l(p) = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \\ 0 \end{bmatrix} \text{ and } r^* = 1$$

we have

$$\phi_{r^*}(s) = (-s_1)^1 (-s_2)^0 (-20/\sqrt{5} x_2) \text{ and } x_2 = 0 \text{ in } D_1.$$

By Theorem 3.1, part 2), our system is controllable on D_1 .

Similar arguments show that the system is controllable on D_2, D_3 , and D_4 . However, we can move from D_1 to D_2 , from D_2 to D_3 , from D_3 to D_4 , and from D_4 to D_1 by using the vector field f . Thus, the system is actually controllable on $M = \mathbb{R}^3 - (x_1\text{-axis})$. Note that the vector space dimension of L_4 is 1 on the x_1 -axis.

In our last example we were able to apply our theory on D_1, D_2, D_3 , and D_4 and make observations concerning $\partial D_1, \partial D_2, \partial D_3, \partial D_4$ in \mathbb{R}^3 to show the system was controllable on $M = \mathbb{R}^3 - (x_1\text{-axis})$. We remark that $\partial D_1 - (x_1\text{-axis})$ consists of the disjoint union of two integral manifolds of g_1 and g_2 , neither of which disconnects M or intersects the lines $x_1 = \pm 2x_2$. The same is true for each of $\partial D_2, \partial D_3$, and ∂D_4 . This illustrates how our theory together with other techniques can yield global solutions. We now give an example in which the control vector is not constant or linear.

Example 3.8: We let $M = \mathbb{R}^2$ and

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_1^2 \\ 0 \end{bmatrix} + u \begin{bmatrix} 0 \\ x_1^2 \end{bmatrix} = f(x(t)) + u(t)g(x(t)).$$

Set $D^+ = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > 0\}$ and $D^- = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 < 0\}$. Obviously, we can move from D^+ to D^- or D^- to D^+ using the vector field f . Thus, we prove that the system is controllable on D^+ , a similar argument being applied on D^- to give us controllability on \mathbb{R}^2 .

The vector fields f and g are linearly dependent in D^+ if $x_2 = 0$, and every integral curve of g contains such a point. Computations yield

$$[f, g] = \begin{bmatrix} -3x_1^2 x_2 \\ 0 \end{bmatrix},$$

$$[[f, g], g] = \begin{bmatrix} 6x_1^2 x_2^2 \\ 0 \end{bmatrix},$$

$$[[[f, g], g], g] = \begin{bmatrix} -6x_1^2 x_2^3 \\ 0 \end{bmatrix}.$$

If we let $l(p) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $x_2 = 0$, and $r^* = 3$, we have

$$\phi_{r^*}(s) = (-s)^3 (-6x_1^2)$$

which satisfies condition 2) of Theorem 3.1 for every $x_1 > 0$. We have a controllable system.

In this paper we have concentrated on the problem of controllability for systems with $(n-1)$ controls on an n -dimensional manifold. An open problem is to provide a similar theory if the number of control vectors is less than $(n-1)$ using the results of [1], [7], and [8].

REFERENCES

- [1] A. Bacciotti and G. Stefani, "The region of attainability of nonlinear systems with unbounded controls," to be published.
- [2] H. Federer, *Geometric Measure Theory*. New York: Springer, 1969.
- [3] H. Hermes, "Controlled stability," *Ann. Mat. Pur. Appl.*, vol. 114, pp. 103-119, 1977.
- [4] R. M. Hirschorn, Global controllability of nonlinear systems, *SIAM J. Contr.*, vol. 14, pp. 700-711, 1976.
- [5] L. R. Hunt, "Global controllability of nonlinear systems in two dimensions," *Math. Syst. Theory*, vol. 13, pp. 361-376, 1980.
- [6] —, "Controllability of nonlinear hypersurface systems," in *Algebraic and Geometric Methods in Linear Systems Theory, AMS Lectures in Applied Mathematics*, vol. 18, C. I. Byrnes and C. F. Martin, Eds. pp. 209-224.
- [7] —, "Control theory for nonlinear systems," in *1979 Int. Symp. Math. Theory Networks and Syst.*, vol. 3, 1979, pp. 339-343.
- [8] —, "Controllability of general nonlinear systems," *Math. Syst. Theory*, vol. 12, pp. 361-370, 1979.
- [9] A. J. Krener, "A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems," *SIAM J. Contr.*, vol. 12, pp. 43-52, 1974.
- [10] R. Stangelaad, "Some aspects of systems and control," Master's thesis, Texas Tech Univ., Lubbock.
- [11] H. Sussman and V. Jurdjevic, "Controllability of nonlinear systems," *J. Differential Equations*, vol. 12, pp. 95-116, 1972.

170

Sufficient Conditions for Controllability

LOUIS R. HUNT, MEMBER, IEEE

Abstract—The problem is to find sufficient conditions for the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^m u_i(t)g_i(x(t)), \quad x(0) = x_0 \in M$$

to be controllable. Here M is a connected C^∞ n -dimensional manifold, f, g_1, \dots, g_m are complete C^∞ vector fields on M , and u_1, \dots, u_m are real-valued controls. If $m = n-1$, $M, f, g_1, \dots, g_{n-1}$ are real-analytic, M is simply connected, and g_1, \dots, g_{n-1} are linearly independent on M , then necessary and sufficient conditions are known. For the case of our C^∞ system with general m , we assume that the space spanned by the Lie algebra L_A generated by f, g_1, \dots, g_m and successive Lie brackets has constant dimension p on M and the algebra L'_A generated by g_1, \dots, g_m and successive Lie brackets has constant dimension $p' \leq p$ on M . If $p' = p$, Chow's Theorem implies controllability for a p -dimensional submanifold of M containing x_0 . If $p' < p$, sufficient conditions are found involving the computation of certain Lie brackets at points where the vector field f is tangent to the integral manifolds of L'_A . Here we assume that every integral manifold of L'_A contains such a point.

I. INTRODUCTION

LET M be a connected C^∞ n -dimensional paracompact manifold, f, g_1, \dots, g_m be complete C^∞ vector fields on M , and u_1, \dots, u_m be unbounded (or unlimited) real-valued controls. We are interested in finding sufficient conditions that the system

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^m u_i(t)g_i(x(t)), \quad x(0) = x_0 \in M \quad (1)$$

be controllable. Let L_A be the Lie algebra generated by f, g_1, \dots, g_m and successive Lie brackets, and suppose that the vector space dimension (i.e., the space spanned by the vector fields in L_A) of L_A on M is p . By Chow's Theorem the set of points in M which are reachable from x_0 are contained in a C^∞ p -dimensional submanifold S of M through the point x_0 . We could develop our controllability theory for this submanifold S , but to conserve notation we assume that $p = n$ and $S = M$.

Thus the supposition is made that the vector space dimension of L_A on M is n at every point of M . Hence, given an arbitrary point x_0 in M , there is an open subset of M which is reachable from x_0 by results from [8] and [10]. In [6] it is shown that this open set can be taken so that its closure contains x_0 . A characterization of the largest open subset of M which is reachable from x_0 , called the region of reachability from x_0 , is given in [1], [6], and [7].

Under the assumption that $m = n-1$, $M, f, g_1, \dots, g_{n-1}$ are real-analytic on M , M is simply connected, and g_1, \dots, g_{n-1} are linearly independent on M , necessary and sufficient conditions are proved in [4] for our system to be controllable on M . These conditions depend on results found in [2] and [5]. It is assumed that the vector space dimension of the Lie algebra L'_A generated by g_1, \dots, g_{n-1} and successive Lie brackets is greater than or equal to $(n-1)$.

With our C^∞ manifold M and general m we suppose that the vector space dimension (i.e., the spanning dimension) of L'_A at every point of M is the constant k . If $k = n$, Chow's Theorem implies controllability of the system on M . If $k < n$, results from [6] and [7] mentioned previously and a theorem due to Hermes [3] implying local controllability along a reference solution are combined to give sufficient controllability conditions.

If every integral manifold N of L'_A in M contains a point x where f is tangent to N (included in this is the possibility that f vanishes at x), then our system is controllable if the following conditions hold for at least one such x in each N . There exist a basis h_1, \dots, h_k of L'_A near x and integers l_1, \dots, l_k such that the space spanned by

$$\{h_1(x), \dots, h_k(x), (adf, h_1)(x), \dots, (adf, h_k)(x), \dots, (ad^{l_1}f, h_1)(x), \dots, (ad^{l_k}f, h_k)(x)\}$$

has dimension n . The vector field \hat{f} is defined in the following manner. Let c_1, \dots, c_k be constants so that $f - c_1h_1 - c_2h_2 - \dots - c_kh_k$ vanishes at x and set

$$\hat{f} = f - \sum_{i=1}^k c_i h_i.$$

Also $(adf, h) = [\hat{f}, h]$, the Lie bracket of \hat{f} and h , $(ad^2\hat{f}, h) \equiv [\hat{f}, [\hat{f}, h]]$, etc. If g_1, \dots, g_m are linearly independent and involutive on M , the above condition is replaced by the space spanned by

$$\{g_1(x), \dots, g_m(x), (adf, g_1)(x), \dots, (adf, g_m)(x), \dots, (ad^{l_1}\hat{f}, g_1)(x), \dots, (ad^{l_m}\hat{f}, g_m)(x)\}$$

is of dimension n . Of course this generalizes the known controllability matrix criterion for linear time-invariant systems.

II. DEFINITIONS

We are interested in the controllability properties of the system (1).

Let $T(M)$ be the tangent bundle to M with fiber $T_x(M)$ for $x \in M$, the tangent space to M at x . For X a C^∞ vector

Manuscript received February 9, 1981; revised July 7, 1981 and November 6, 1981. This work was supported in part by the Joint Services Electronics Program under ONR Contract N00014-76-C1136.

The author is with NASA Ames Research Center, MS 210-3, Moffett Field, CA 94035, on leave from the Department of Mathematics, Texas Tech University, Lubbock, TX 79409.

field on M , α is an integral curve of X if α is a C^∞ mapping from an interval $I \subset \mathbb{R}$ into M such that $d\alpha(t)/dt = X(\alpha(t))$ for all $t \in I$. If T is a subset of $T(M)$, an *integral curve of T* is a mapping α' from a real interval $[t, t']$ into M so that there exist $t = t_0 < t_1 < t_2 < \dots < t_j = t'$ and vector fields X_1, \dots, X_j in T with the restriction of α' to $[t_{i-1}, t_i]$ being an integral curve of X_i for each $i = 1, 2, \dots, j$. A connected submanifold N of M is an *integral manifold* of vector fields X_1, \dots, X_k if $T_i(N)$ is the space spanned by X_1, \dots, X_k at y for each $y \in N$.

The subset T of M we consider is the one given by the vector fields in our system (1). A point $x \in M$ is *reachable from $x_0 \in M$* if there is an integral curve α' of T and some time $t \geq 0$ in the interval for α' such that $\alpha'(0) = x_0$ and $\alpha'(t) = x$. A subset A of M is *reachable from x_0* if every point $x \in A$ is reachable from x_0 . If an open subset of M is reachable from x_0 , the largest such open set is called the *region of reachability from x_0* . If the system is controllable from every x_0 in M , then it is *controllable*. Let $\varphi(t, x_0)$ be the solution of (1) at time t with $\varphi(0, x_0) = x_0$ and which corresponds to all controls u_i being 0. The system (1) is *locally controllable along φ at time $t \geq 0$* if all points in some n -dimensional neighborhood of $\varphi(t, x_0)$ can be reached at time t by solutions of (1) initiating from x_0 .

Let O be an open subset of M and take $x \in \partial O$. Then f points in the direction of \bar{O} (or toward \bar{O}) at x if there is an open neighborhood W of x in M so that the integral curve of f starting at x and intersected with W is contained in \bar{O} , the closure of O in M . If this is true for all $x \in \partial O$, then f points in the direction of \bar{O} on ∂O . We define f pointing in the direction of O by replacing \bar{O} everywhere with O .

If X and Y are C^∞ vector fields on M , we define the *Lie bracket* of X and Y by

$$[X, Y] = \frac{\partial Y}{\partial x} X - \frac{\partial X}{\partial x} Y$$

where $\partial Y/\partial x$ and $\partial X/\partial x$ are Jacobian matrices. Higher order Lie brackets such as $[X, [X, Y]]$, $[Y, [X, Y]]$, \dots can be defined. We let $(ad^0 X, Y) = [X, Y]$ and inductively $(ad^{i-1} X, Y) = [X, (ad^i X, Y)]$. By L_A and L'_A we denote the respective Lie algebras generated by f, g_1, \dots, g_m and successive brackets and g_1, \dots, g_m and successive brackets where f, g_1, \dots, g_m are defined in (1). The set of vector fields g_1, \dots, g_m is called *involutive* if there exist functions γ_{ijk} such that

$$[g_i, g_j](x) = \sum_{k=1}^m \gamma_{ijk}(x) g_k(x)$$

for all $i, j, 1 \leq i, j \leq m, i \neq j$.

III. MAIN RESULTS

To prove our controllability results we combine the local controllability along a trajectory work of Hermes [3] with the local to global theory found in [7]. The first theorem is due to Hermes.

Theorem 3.1. Let $\varphi(t, x_0)$ be a solution of (1) corresponding to all $u_i = 0$. A sufficient condition that the

system (1) be locally controllable along φ at time $t \geq 0$ is that there exist integers l_1, \dots, l_m such that the space spanned by $\{g_1(x_0), \dots, g_m(x_0), (ad^{l_1} f, g_1)(x_0), \dots, (ad^{l_m} f, g_m)(x_0), \dots, (ad^{l_1} f, g_1)(x_0), \dots, (ad^{l_m} f, g_m)(x_0)\}$ is of dimension n .

If x_0 is a critical point (or equilibrium point) for f , then an open neighborhood of x_0 in M can be reached by trajectories of the system (1) starting at x_0 .

Our next theorem was first proved in [6] and [7] and improved by Bacciotti and Stefani [1]. We assume that the vector space dimensions of L_A and L'_A are constant on M .

Theorem 3.2. Let U be the smallest open subset of M containing x_0 in its closure such that ∂U contains the C^∞ k -dimensional integral manifolds of L'_A that intersect it. If f points in the direction of \bar{U} on ∂U , then

$$U \subset \{\text{region of reachability from } x_0\} \subset \text{interior of } \bar{U}.$$

Conversely, if U is the region of reachability from x_0 , then ∂U contains the k -dimensional integral manifolds of L'_A which intersect it, f points toward \bar{U} on ∂U , and U contains x_0 in its closure.

Let \bar{N} be any connected $(n-1)$ -dimensional manifold (not necessarily differentiable) consisting of the union of an integral manifold N with other integral manifolds of L'_A near N . If every integral manifold N of L'_A in M contains a point x which has the following property, then the system (1) is controllable by Theorem 3.2. There is an open neighborhood W of x in M with \bar{N} separating W into disjoint nonempty open sets O_1 and O_2 (with \bar{N} being the common boundary of O_1 and O_2 in W) so that f points toward O_1 at some $x_1 \in \bar{N} \cap W$ and f points toward O_2 at some $x_2 \in \bar{N} \cap W$. However, if f has a critical point at x , if there is a neighborhood W of x with N, \bar{N}, O_1 , and O_2 as above, and if f points toward \bar{O}_1 (or \bar{O}_2) at all points of $\bar{N} \cap W$, then the system is not locally controllable along φ .

In our next theorem the vector field \hat{f} is defined in the following way. If x is a critical point of f , then let $\hat{f} = f$. If x is a point where f is nonzero and is tangent to the integral manifold N of L'_A through x , then for h_1, \dots, h_k a basis for L'_A near x , there exist constants c_1, \dots, c_k so that $\hat{f} = c_1 h_1 + c_2 h_2 + \dots + c_k h_k$ is 0 at x . We set

$$\hat{f} = f - \sum_{i=1}^k c_i h_i.$$

Theorem 3.3. Suppose every integral manifold N of L'_A in M contains a point x where \hat{f} is tangent to N and there exist a basis h_1, \dots, h_k of L'_A near x and integers l_1, \dots, l_k such that the space spanned by

$$\{h_1(x), \dots, h_k(x), (ad^{l_1} \hat{f}, h_1)(x), \dots, (ad^{l_k} \hat{f}, h_k)(x), \dots, (ad^{l_1} \hat{f}, h_1)(x), \dots, (ad^{l_k} \hat{f}, h_k)(x)\}$$

has dimension n . Then the system (1) is controllable.

Proof: We examine the system

$$\dot{x}(t) = f(x(t)) - \sum_{i=1}^k u_i(t) h_i(x(t)) \quad (2)$$

where the u_i are unlimited controls. With \hat{f} as defined earlier the point x is a critical point for the system

$$\dot{x}(t) = \hat{f}(x(t)) + \sum_{i=1}^k u_i(t) h_i(x(t)). \quad (3)$$

From Theorem 3.1 we deduce that the system (3) is locally controllable along the trajectory with all $u_i = 0$. By the comment after that result, an open neighborhood of x in M is reachable from x . Given an open neighborhood W of x and a $(n-1)$ -dimensional manifold \tilde{N} separating W into open sets O_1 and O_2 as before, we must have that \hat{f} points toward O_1 at some $x_1 \in \tilde{N} \cap W$ and toward O_2 at some $x_2 \in \tilde{N} \cap W$ by a previous remark. Since f and \hat{f} differ by vector fields in the tangent bundle of L'_A near x , the same is true for f itself. From statements in the paragraph preceding this theorem, we have that our system (1) is controllable on M .

Remark. Given any point x_0 in M , the research of Sussmann and Jurdjević [10] and Krener [8] implies that there is an open set in M which is reachable from x_0 (a local result). Hermes [3] shows that under the assumptions of Theorem 3.1 for a critical point x_0 of f , this open set is actually an open neighborhood of x_0 (again a local result). The important ideas in Theorem 3.3 are

(a) that the points x where f is tangent to the integral manifolds of L'_A can be treated by Hermes' theory to yield local controllability, and

(b) the local to global capabilities of Theorem 3.2 are applied to prove a result on global controllability which depends on examining certain Lie brackets at only those points where f is tangent to the integral manifolds.

In general, global theorems are difficult to prove, and proceeding from the local to the global is quite a problem.

We have the following corollary to Theorem 3.3.

Corollary 3.4. Let g_1, \dots, g_m be involutive on M . Suppose every integral manifold N of g_1, \dots, g_m in M contains a point x where f is tangent to N and there exist integers l_1, \dots, l_m so that the space spanned by $\{g_1(x), \dots, g_m(x), (ad^{l_1} f, g_1)(x), \dots, (ad^{l_m} f, g_m)(x), \dots, (ad^{l_1} f, g_1)(x), \dots, (ad^{l_m} f, g_m)(x)\}$ has dimension n . Then the system (1) is controllable.

As an application of our theory we present the following example suggested by Meyer [9].

The state space is four dimensional but 2 axis, so that the state $x = (y_1, y_2)$ and both y_1 and y_2 are two dimensional with

$$y_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad y_2 = \begin{bmatrix} x_3 \\ x_4 \end{bmatrix}.$$

The vector y_1 is the position vector in two dimensions and the velocity \dot{y}_1 is generated by y_2 through two-dimensional rotation, given by the matrix

$$E = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

where the angle θ is a function of the Euclidean distance $\|y_1\|$ from the origin in \mathbb{R}^2 . The control u is 2 axis and the

state equation is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} (\cos \theta)x_3 + (\sin \theta)x_4 \\ (-\sin \theta)x_3 + (\cos \theta)x_4 \\ 0 \\ 0 \end{bmatrix} + u_1 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \\ = f(x) + u_1 g_1(x) + u_2 g_2(x)$$

with $M = \mathbb{R}^4$. Now g_1 and g_2 are involutive on \mathbb{R}^4 , and f is tangent to the integral manifolds of g_1 and g_2 if $x_3 = x_4 = 0$. Certainly every integral manifold N of g_1 and g_2 contains such a point and $\hat{f} = f$ at all such points.

Computing Lie brackets we find that

$$(adf, g_1) = - \begin{bmatrix} \cos \theta \\ -\sin \theta \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad (adf, g_2) = - \begin{bmatrix} \sin \theta \\ \cos \theta \\ 0 \\ 0 \end{bmatrix}.$$

Thus $g_1, g_2, (adf, g_1)$, and (adf, g_2) are linearly independent when $x_3 = x_4 = 0$, and by Corollary 3.4 our system is controllable.

In the following example the control vectors are not involutive and Theorem 3.3 must be applied.

Let

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \begin{bmatrix} \sin x_4 \\ \sin x_5 \\ x_3 \\ x_2 \\ x_1 \end{bmatrix} + u_1 \begin{bmatrix} 0 \\ 0 \\ x_4 \\ x_5 \\ x_3 \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \\ = f(x(t)) + \sum_{i=1}^2 u_i(t) g_i(x(t))$$

where $M = \mathbb{R}^5$.

Computing Lie brackets we find that

$$[g_1, g_2] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad [g_1, [g_1, g_2]] = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

implying that the vector space dimension of L'_A on M is 3. An appropriate basis for L'_A is

$$h_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad h_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad h_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

The vector field f is tangent to the integral manifolds of L'_A if and only if x_4 and x_5 are integer multiples of π . Certainly every integral manifold N of L'_A contains infinitely many such points.

We have

$$[f, h_2] = - \begin{bmatrix} \cos x_4 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad [f, h_1] = - \begin{bmatrix} 0 \\ \cos x_5 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Since the computations with f and \hat{f} yield the same results for this example, we find that $\{h_1(x), h_2(x), h_3(x), (adf, h_1)(x), (adf, h_2)(x)\}$ span a five-dimensional space when x_4 and x_5 are integer multiples of π . Thus our system is controllable.

A much more difficult problem is encountered if there exists an integral manifold N of L'_A which does not contain a point x where f is tangent to N .

REFERENCES

- [1] A. Bacciotti and G. Stefani, "The region of attainability of nonlinear systems with unbounded controls," in preparation.
- [2] H. Hermes, "Controlled stability," *Ann. Mat. Pur. Appl.*, vol. 114, pp. 103-119, 1977.
- [3] —, "On local and global controllability," *SIAM J. Contr.*, vol. 12, pp. 252-261, 1974.
- [4] L. R. Hunt, "n-dimensional controllability with $(n-1)$ controls," *IEEE Trans. Automat. Contr.*, to appear.
- [5] —, "Controllability of nonlinear hypersurface systems," in *Algebraic and Geometric Methods in Linear Systems Theory, Lectures in Applied Mathematics*, vol. 18, C. I. Byrnes and C. F. Martin, Eds., pp. 209-224, 1980.
- [6] —, "Control theory for nonlinear systems," in *Int. Symp. Mathematical Theory of Networks and Systems*, (1979), pp. 339-343, 1979.
- [7] —, "Controllability of general nonlinear systems," *Math. Systems Theory*, vol. 12, pp. 361-370, 1979.
- [8] A. J. Krener, "A generalization of Chow's Theorem and the bang-bang theorem to nonlinear control problems," *SIAM J. Contr.*, vol. 12, pp. 43-52, 1974.
- [9] G. Meyer and L. Gicolanu, "Application of nonlinear system inverses to automatic flight control design-system concepts and flight evaluations," in *AGARDograph on Theory and Applications of Optimal Control in Aerospace Systems*, P. Kant, Ed., 1980.
- [10] H. Sussmann and V. Jurdjevic, "Controllability of nonlinear systems," *J. Differential Equation*, vol. 12, pp. 95-116, 1972.

+



Louis R. Hunt (M'79) was born in Shreveport, LA, on December 5, 1942. He received the B. S. degree in 1964 from Baylor University and the Ph.D. degree in 1970 from Rice University, both in mathematics.

He is Professor of Mathematics at Texas Tech University, but is currently on leave at NASA Ames Research Center, Moffett Field, CA. Dr. Hunt's research interests are nonlinear systems and control, several complex variables, and partial differential equations. At NASA Ames, he is working with a group concerned with automatic flight control of aircraft.

He is a member of the American Mathematical Society and the Society for Industrial and Applied Mathematics.

CONTROL OF NONLINEAR TIME-VARYING SYSTEMS

L. R. Hunt* and Renjeng Su†

Ames Research Center, NASA, Mail Stop 210-3,
Moffett Field, California 94035

Abstract

Consider the time-varying nonlinear system of the form $\dot{x}(t) = f(x,t) + \sum_{i=1}^m u_i(t)g_i(x,t)$, with f, g_1, \dots, g_m being \mathcal{C}^∞ vector fields on \mathbb{R}^{n+1} . We give necessary and sufficient conditions for this system to be transformable to a time-invariant controllable linear system. In order to control the nonlinear system, we map to the linear system, choose a desired control there, and return to the nonlinear system by the inverse of the transformation.

I. Introduction

Suppose we wish to control a nonlinear system of the form

$$\dot{x}(t) = f(x,t) + \sum_{i=1}^m u_i(t)g_i(x,t) \quad (1)$$

where f, g_1, \dots, g_m are \mathcal{C}^∞ complete vector fields on $\mathbb{R}^n \times \mathbb{R}$, $f(0,t) = 0$ for all t , and g_1, g_2, \dots, g_m are linearly independent. If we can find a nonsingular one-to-one \mathcal{C}^∞ transformation that maps this nonlinear time-varying system to a controllable linear time-invariant system, then we can use the control theory for the linear system to control the nonlinear system. In this paper we consider only the local case, where the transformation maps a neighborhood of the origin in \mathbb{R}^{n+m} to a neighborhood of the origin in \mathbb{R}^{n+m} for each t .

Since every controllable linear system has as invariants the Kronecker indices and with these a Brunovsky [1] canonical form, in order to map a nonlinear system to a controllable linear system we assume that the linear system is in Brunovsky form. From this point of view we may take an arbitrary set of Kronecker indices $\kappa_1, \kappa_2, \dots, \kappa_m$, with $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_m$, and use the associated linear system as our target.

*Author on leave from Department of Mathematics, Texas Tech University, Lubbock, TX 79409. Research supported by NASA Ames Research Center under the IPA Program and the Joint Services Electronics Program at Texas Tech University under ONR Contract N00014-76-C-1126.

†Research supported by NASA Ames Research Center and the National Research Council.

We indicate the type of mappings of interest to us. A \mathcal{C}^∞ transformation $T = (T_1, T_2, \dots, T_{n+m})$ maps an open set in $\mathbb{R}^{n+m+1}((x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m, t)$ space) containing the origin in \mathbb{R}^{n+m} onto an open set in $\mathbb{R}^{n+m}(T_1, T_2, \dots, T_n, T_{n+1}, \dots, T_{n+m})$ space) containing the origin in \mathbb{R}^{n+m} such that the following properties hold:

- i) $T(0,t) = 0$ for all t ;
- ii) T_1, T_2, \dots, T_n are functions of x_1, x_2, \dots, x_n and t only and have a nonsingular Jacobian matrix in some open neighborhood of the origin in \mathbb{R}^n for each fixed t ;
- iii) $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ are functions of $x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m, t$ and for fixed (x_1, x_2, \dots, x_n) near the origin, the $m \times m$ Jacobian matrix of $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ with respect to u_1, u_2, \dots, u_m is nonsingular, again, for every t ;
- iv) T_1, T_2, \dots, T_n are the state variables and $T_{n+1}, T_{n+2}, \dots, T_{n+m}$ are the controls for a linear time-invariant system in the appropriate Brunovsky canonical form; and
- v) for each fixed t , $T = (T_1, T_2, \dots, T_{n+m})$ is a one-to-one map of an open neighborhood of the origin in $(x_1, x_2, \dots, x_n, u_1, u_2, \dots, u_m)$ space onto an open neighborhood of the origin in $(T_1, T_2, \dots, T_n, T_{n+1}, T_{n+2}, \dots, T_{n+m})$ space. Here, u_1, \dots, u_m and T_{n+1}, \dots, T_{n+m} can be as large as we wish.

In other words, we want a local diffeomorphism T (for fixed t) which maps system (1) to our linear system. Since our work will be in a neighborhood of the origin, we find it unnecessary to name specific sets; we suppose that all assumptions, conditions, and results hold in an open set in the appropriate Euclidean space that contains the origin. This theory can be combined with the global inverse-function theorems as in [2] to produce global results.

If our system (1) is autonomous, then the following result from [3] gives necessary and sufficient conditions to map to the Brunovsky form.

with indices $\kappa_1, \kappa_2, \dots, \kappa_m$. With a possible reordering of the g 's, we can transform our nonlinear autonomous system to the linear system with indices $\kappa_1, \kappa_2, \dots, \kappa_m$ if and only if

a) the set $C = \{g_1, [f, g_1], \dots, (ad^{\kappa_1-1} f, g_1), g_2, [f, g_2], \dots, (ad^{\kappa_2-1} f, g_2), \dots, g_m, [f, g_m], \dots, (ad^{\kappa_m-1} f, g_m)\}$ spans an n -dimensional space;

b) the sets $C_j = \{g_j, [f, g_j], \dots, (ad^{\kappa_j-1} f, g_j), g_2, [f, g_2], \dots, (ad^{\kappa_2-1} f, g_2), \dots, g_m, [f, g_m], \dots, (ad^{\kappa_m-1} f, g_m)\}$ are involutive for $j = 1, 2, \dots, m$; and

c) the span of each C_j is equal to the span of $C_j \cap C$.

Here $[f, g_1]$ is the Lie bracket, $(ad^0 f, g_1) = g_1$; $(ad^1 f, g_1) = [f, g_1]$; $(ad^2 f, g_1) = [f, [f, g_1]]$; etc.

It is the purpose of this paper to generalize this result to our nonlinear time-varying system (1). We give an example of a constructive proof of the transformation. Details, proofs, and a description of how to apply the theory in practice will appear elsewhere.

Historically, the problem of transforming a periodic time-varying linear system to a linear time-invariant system is due to Floquet and Liapunov. Brunovsky [1] gave necessary and sufficient conditions for a linear time-varying system to be "equivalent" to a controllable time-invariant one. Meyer and Cicolani in [4] and [5] present conditions for their nonlinear time-varying block triangular systems to be transformed to a linear time-invariant system. Their block triangular systems can be put in the form (1). The second author in [6] has given a talk on the single-input case.

For transformations of autonomous nonlinear systems to linear systems we refer to the work of Krener [7], Brockett [8], Jakubczyk and Respondek [9], and the authors ([2], [4], [6], and [10]).

In section II, we give definitions and preliminaries, and section III contains our main result on transforming nonlinear systems.

II. Definitions

If f and g are \mathcal{C}^∞ vector fields on a manifold M (\mathbb{R}^n in our theory), the Lie bracket of f and g is

$$[f, g] = \frac{\partial g}{\partial x} f - \frac{\partial f}{\partial x} g \quad (2)$$

where $\partial g / \partial x$ and $\partial f / \partial x$ denote Jacobian matrices. We define

$$\begin{aligned} (ad^0 f, g) &= g \\ (ad^1 f, g) &= [f, g] \\ (ad^2 f, g) &= [f, [f, g]] \\ &\vdots \\ (ad^k f, g) &= [f, (ad^{k-1} f, g)] \end{aligned} \quad (3)$$

Given vector fields f_1, f_2, \dots, f_r on \mathbb{R}^n , we say that this set is involutive if there exist \mathcal{C}^∞ functions γ_{ijk} with

$$[f_i, f_j](x) = \sum_{k=1}^r \gamma_{ijk}(x) f_k(x), \quad 1 \leq i, j \leq r, i \neq j.$$

If these vector fields are involutive and $x_0 \in \mathbb{R}^n$, then there is a unique r -dimensional \mathcal{C}^∞ manifold S containing x_0 so that the tangent space to S at each $x \in S$ is the space spanned by $f_1(x), f_2(x), \dots, f_r(x)$. That is, S is the unique integral manifold of f_1, f_2, \dots, f_r through x_0 . This is, of course, due to the famous theorem of Frobenius.

For a \mathcal{C}^∞ vector field f and a \mathcal{C}^∞ time-varying vector field g , we define

$$\begin{aligned} (r^0 f, g) &= g \\ (r^1 f, g) &= (ad^1 f, g) + \frac{\partial g}{\partial t} \\ &\vdots \\ (r^k f, g) &= (r^1 f, (r^{k-1} f, g)) \end{aligned} \quad (4)$$

If h is a \mathcal{C}^∞ function, we define the Lie derivative of h with respect to f as

$$L_f(h) = \langle dh, f \rangle \quad (5)$$

where $\langle \cdot, \cdot \rangle$ denotes the duality between one forms and vector fields. Similarly, we let

$$\begin{aligned} L_f^0(h) &= h \\ L_f^1(h) &= L_f(h) \\ &\vdots \\ L_f^k(h) &= L_f^1(L_f^{k-1}(h)) \end{aligned} \quad (6)$$

For a \mathcal{C}^∞ one form ω , we have

$$L_f(\omega) = \left(\frac{\partial \omega^*}{\partial x} f \right)^* + \omega \frac{\partial f}{\partial x} = d\langle \omega, f \rangle \quad (7)$$

where $*$ denotes the transpose and $\partial \omega^* / \partial x$ and $\partial f / \partial x$ are Jacobian matrices. For this derivative, we have

$$\begin{aligned} L_f^0(\omega) &= \omega \\ L_f^1(\omega) &= L_f(\omega) \\ &\vdots \\ L_f^k(\omega) &= L_f^1(L_f^{k-1}(\omega)) \end{aligned} \quad (8)$$

An important formula relating some of these Lie derivatives is

$$L_f^1 \langle \omega, g \rangle = \langle L_f^1(\omega), g \rangle + \langle \omega, [f, g] \rangle \quad (9)$$

As noted before, we wish to transform system (1) to the system

$$\dot{y} = Ay + vB \quad (10)$$

where we have the Brunovsky canonical form associated with $\kappa_1, \kappa_2, \dots, \kappa_m$. Here we have set $y = (y_1, y_2, \dots, y_n) = (T_1, T_2, \dots, T_n)$ and $v = (v_1, v_2, \dots, v_m) = (T_{n+1}, T_{n+2}, \dots, T_{n+m})$. The matrix A is

$$\begin{array}{c} \left. \begin{array}{c} \kappa_1 \\ \kappa_2 \\ \vdots \\ \kappa_m \end{array} \right\} \left[\begin{array}{cccc|cccc|cccc|cccc} 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & & & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & & & & & & \\ \vdots & & & \ddots & & & & & \ddots & & & & & & & \\ & & & & 1 & & & & & & & & & & & \\ 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 & & & 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & 0 & 0 & 1 & 0 & \dots & 0 & & 0 & 0 & \dots & \dots & 0 \\ \vdots & & & & & 0 & 0 & 1 & 0 & \dots & 0 & & & & & \\ & & & & & & & & & 1 & & & & & & \\ 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 & & & 0 & 0 & \dots & \dots & 0 \end{array} \right] \quad (11)$$

$$\left. \begin{array}{c} \vdots \\ \vdots \\ \vdots \end{array} \right\} \left[\begin{array}{cccc|cccc|cccc|cccc} 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 & & & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & & & & & & & & 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & & & & & & & & & & & & & \\ & & & & & & & & & & & & & & & & 1 \\ 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 & & & 0 & 0 & \dots & \dots & 0 \end{array} \right]$$

and matrix B is

$$B = \begin{bmatrix} \begin{matrix} \kappa_1 \\ \vdots \\ \kappa_m \end{matrix} \left\{ \begin{array}{cccccc} 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & 0 & \dots & \dots & \dots & 0 \\ \hline 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 1 & 0 & \dots & \dots & 0 \\ \hline \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \hline 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & \dots & 1 \end{array} \right\} \end{bmatrix} \quad (12)$$

III. Partial Differential Equations

The problem of transforming system (1) to system (10) reduces to the study of a system of partial differential equations. We let $\sigma_1 = \kappa_1$, $\sigma_2 = \kappa_1 + \kappa_2$, ..., $\sigma_m = \kappa_1 + \kappa_2 + \dots + \kappa_m = n$; note that the equations in the following lemma are derived in [3] for the autonomous case.

Lemma 3.1.

The system (1) is transformable to the linear system (10) if, and only if, we can solve the equations

$$\begin{aligned} \langle dT_i, g_i \rangle &= 0, \quad i = 1, 2, \dots, \\ \sigma_1 - 1, \sigma_1 + 1, \dots, \\ \sigma_2 - 1, \sigma_2 + 1, \dots, \\ \sigma_{m-1} - 1, \sigma_{m-1} + 1, \dots, \\ n - 1 \text{ and } i = 1, 2, \dots, m \\ \langle dT_i, f \rangle + \frac{\partial T_i}{\partial t} &= T_{i+1}, \quad i = 1, 2, \dots, \\ \sigma_1 - 1, \sigma_1 + 1, \dots, \\ \sigma_2 - 1, \sigma_2 + 1, \dots, \\ \sigma_{m-1} - 1, \sigma_{m-1} + 1, \dots, n - 1 \end{aligned}$$

(13)

$$\left\langle dT_{\sigma_1}, f + \sum_{i=1}^m u_i g_i \right\rangle + \frac{\partial T_{\sigma_1}}{\partial t} = T_{n+1}$$

$$\left\langle dT_{\sigma_2}, f + \sum_{i=1}^m u_i g_i \right\rangle + \frac{\partial T_{\sigma_2}}{\partial t} = T_{n+2}$$

$$\left\langle dT_{\sigma_m}, f + \sum_{i=1}^m u_i g_i \right\rangle + \frac{\partial T_{\sigma_m}}{\partial t} = T_{n+m}$$

(13)
(concluded)

with the matrix

$$\begin{bmatrix} \langle dT_{\sigma_1}, g_1 \rangle & \langle dT_{\sigma_1}, g_2 \rangle & \dots & \langle dT_{\sigma_1}, g_m \rangle \\ \langle dT_{\sigma_2}, g_1 \rangle & \langle dT_{\sigma_2}, g_2 \rangle & \dots & \langle dT_{\sigma_2}, g_m \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle dT_{\sigma_m}, g_1 \rangle & \langle dT_{\sigma_m}, g_2 \rangle & \dots & \langle dT_{\sigma_m}, g_m \rangle \end{bmatrix} \quad (14)$$

being nonsingular.

Using the Leibnitz Rule (9) repeatedly, our necessary and sufficient conditions for the existence of a transformation are finding functions $T_1, T_{\sigma_1+1}, \dots, T_{\sigma_{m-1}+1}$ satisfying

$$\langle dT_1, (r^j f, g_i) \rangle = 0, \quad j = 0, 1, \dots, \kappa_1 - 2$$

and $i = 1, 2, \dots, m$

$$\langle dT_{\sigma_1+1}, (r^j f, g_i) \rangle = 0, \quad j = 0, 1, \dots, \kappa_2 - 2$$

and $i = 1, 2, \dots, m$

$$\langle dT_{\sigma_{m-1}+1}, (r^j f, g_i) \rangle = 0, \quad j = 0, 1, \dots, \kappa_m - 2$$

and $i = 1, 2, \dots, m$

(15)

such that the matrix

$$\begin{bmatrix} \langle dT_1, (r^{\kappa_1-1} f, g_1) \rangle & \dots & \langle dT_1, (r^{\kappa_1-1} f, g_m) \rangle \\ \langle dT_{\sigma_1+1}, (r^{\kappa_2-1} f, g_1) \rangle & \dots & \langle dT_{\sigma_1+1}, (r^{\kappa_2-1} f, g_m) \rangle \\ \vdots & \ddots & \vdots \\ \langle dT_{\sigma_{m-1}+1}, (r^{\kappa_m-1} f, g_1) \rangle & \dots & \langle dT_{\sigma_{m-1}+1}, (r^{\kappa_m-1} f, g_m) \rangle \end{bmatrix}$$

(16)

is nonsingular. This leads to our main result.

Theorem 3.2.

The time-varying nonlinear system (1) is transformable to the time-invariant linear system (10) if, and only if,

a) the set $C = \{g_1, (T^1 f, g_1), \dots, (T^{K_1-1} f, g_1), g_2, (T^1 f, g_2), \dots, (T^{K_2-1} f, g_2), \dots, g_m, (T^1 f, g_m), \dots, (T^{K_m-1} f, g_m)\}$ spans an n -dimensional space for each t ;

b) the sets $C_j = \{g_1, (T^1 f, g_1), \dots, (T^{K_j-1} f, g_1), g_2, (T^1 f, g_2), \dots, (T^{K_j-1} f, g_2), \dots, g_m, (T^1 f, g_m), \dots, (T^{K_j-1} f, g_m)\}$ are involutive for each t and $j = 1, 2, \dots, m$; and

c) the span of each C_j is equal to the span of $C_j \cap C$ again for every fixed t .

We show a constructive proof of a solution to (15) and (16) under the hypotheses of Theorem 3.2 and the assumption that $n = 5$, $K_1 = 3$, and $K_2 = 2$. We introduce real parameters s_1, s_2, s_3, s_4, s_5 .

Let $x(s_1)$ be the solution to the system

$$\frac{dx}{ds_1} = (T^2 f, g_1) \quad (17)$$

satisfying $x(0) = (0, 0, \dots, 0)$, the origin in R^5 . Next, we denote by $x(s_1, s_2)$ the solution of

$$\frac{dx}{ds_2} = (T^1 f, g_1) \quad (18)$$

with $x(s_1, 0) = x(s_1)$. The 5-tuple $x(s_1, s_2, s_3)$ is the solution to

$$\frac{dx}{ds_3} = (T^1 f, g_2) \quad (19)$$

with $x(s_1, s_2, 0) = x(s_1, s_2)$. Continuing, we solve in order

$$\frac{dx}{ds_4} = g_1 \quad \text{and} \quad \frac{dx}{ds_5} = g_2 \quad (20)$$

to find a function $x(s_1, s_2, s_3, s_4, s_5)$. From the hypotheses of the theorem, we conclude that the Jacobian matrix of the function with respect to s_1, s_2, s_3, s_4, s_5 is nonsingular at the origin. By the inverse function theorem, we can solve for s_1, s_2, s_3, s_4, s_5 as functions of x_1, x_2, x_3, x_4, x_5 . We note that $x(s_1, s_2, s_3, s_4, s_5)$ is also a function of t , and this inversion is done for each fixed t .

Thus, if we can choose, in the s_1, s_2, s_3, s_4, s_5 space, solutions T_1 and T_4 to

$$\begin{aligned} \langle dT_1, (T^j f, g_1) \rangle &= 0, \quad j = 0, 1 \quad \text{and} \quad i = 1, 2 \\ \langle dT_4, (T^j f, g_1) \rangle &= 0, \quad j = 0 \quad \text{and} \quad i = 1, 2 \end{aligned} \quad (21)$$

then we have a transformation (for our choices the matrix (16) will be nonsingular). We let $T_1 = s_1$ and $T_4 = s_3$. Since $C_1 \cap C$ is involutive for fixed t , we get an integral manifold of $C_1 \cap C$

as we let s_2, s_3, s_4 and s_5 vary (with s_1 fixed). Since $T_1 = s_1$ is constant on each integral manifold, T_1 does satisfy

$$\langle dT_1, (T^j f, g_1) \rangle = 0, \quad j = 0, 1 \quad \text{and} \quad i = 1, 2$$

Also, since $C_2 \cap C$ is involutive with t fixed, letting s_4 and s_5 vary (with s_1, s_2, s_3 fixed), we find an integral manifold of $C_2 \cap C$. Our choice $T_4 = s_3$ is constant on each such integral manifold and

$$\langle dT_4, g_1 \rangle = 0 \quad \text{for} \quad i = 1, 2$$

Example 3.3.

Consider the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \begin{bmatrix} \sin x_2 \\ \sin x_3 \\ 0 \\ x_5 + x_4^5 \\ 0 \end{bmatrix} + u_1 \begin{bmatrix} 0 \\ 0 \\ 1+t \\ 0 \\ 0 \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad (22)$$

on $U \times V \subseteq R^5 \times R$, where

$$U = \{(x_1, x_2, x_3, x_4, x_5) : -\frac{\pi}{2} < x_2, x_3 < \frac{\pi}{2}\}$$

$$V = \{t : -\frac{1}{2} < t\}.$$

Computing, we have

$$\begin{aligned} [f, g_1] &= \begin{bmatrix} 0 \\ -(1+t)\cos x_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \\ (ad^2 f, g_1) &= \begin{bmatrix} \cos x_2 \cos x_3 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad [f, g_2] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}. \end{aligned} \quad (23)$$

Thus

$$(T^1 f, g_1) = \begin{bmatrix} 0 \\ -(1+t)\cos x_3 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad (24)$$

$$(f^2, g_1) = \begin{bmatrix} (1+\tau)(\cos x_2 \cos x_3) \\ -2 \cos x_2 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad (f^1, g_2) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 0 \end{bmatrix} \quad (24)$$

(concluded)

Checking, we find that the hypotheses of Theorem 3.2 are satisfied in a neighborhood of the origin. Solving

$$\begin{aligned} \frac{dx}{ds_1} &= (f^1, g_1) \\ \frac{dx}{ds_2} &= (f^1, g_2) \\ \frac{dx}{ds_3} &= (f^1, g_3) \\ \frac{dx}{ds_4} &= g_4 \\ \frac{dx}{ds_5} &= g_5 \end{aligned} \quad (25)$$

in order, we find

$$\begin{aligned} x_1 &= (1+\tau) \frac{\sin(2s_1)}{2} \\ x_2 &= -(1+\tau)(\sin s_2) - 2s_1 \\ x_3 &= (1+\tau)s_4 + s_2 \\ x_4 &= -s_4 \\ x_5 &= s_5 \end{aligned}$$

Thus, we have

$$\begin{aligned} s_1 &= \frac{1}{2} \arcsin \frac{2x_1}{1+\tau} \\ s_2 &= \arcsin \left(\frac{x_2 + \arcsin \frac{2x_1}{1+\tau}}{-(1+\tau)} \right) \\ s_4 &= -x_4 \\ s_5 &= \frac{x_3 - \arcsin \left(\frac{x_2 + \arcsin \frac{2x_1}{1+\tau}}{-(1+\tau)} \right)}{1+\tau} \\ s_3 &= x_5 \end{aligned}$$

We then define

$$\begin{aligned} T_1 &= \frac{1}{2} \arcsin \left(\frac{2x_1}{1+\tau} \right) \\ T_4 &= -x_4 \end{aligned}$$

The functions T_2, T_3, T_5, T_6 , and T_7 can be found from equations (13).

IV. Conclusions

We have mentioned results that classified those nonlinear time-varying systems like system (1), which can be transformed to controllable linear time-variant systems. We also outlined a method of constructing such a transformation in the case $\kappa_1 = 3$ and $\kappa_2 = 2$ on R^5 .

References

1. P. Brunovsky, "A classification of linear non-controllable systems," Kibernetika (Prague), Vol. 6, pp. 173-188, 1970.
2. L. R. Hunt and R. Su, "Global transformations of nonlinear systems," submitted to IEEE Transactions on Automatic Control.
3. L. R. Hunt and R. Su, "Multi-input nonlinear systems," submitted to SIAM J. Control.
4. G. Meyer and L. Cicolani, "A formal structure for advanced automatic flight control systems," NASA TN D-7940, 1975.
5. G. Meyer and L. Cicolani, "Applications of nonlinear system inverses to automatic flight control design-system concepts and flight evaluation," AGARDograph on Theory and Applications of Optimal Control in Aerospace Systems, P. Kant, ed., 1980.
6. L. R. Hunt and R. Su, "Linear equivalents of nonlinear time-varying systems," 1980 International Symposium on the Mathematical Theory of Networks and Systems, pp. 119-123.
7. A. J. Krener, "On the equivalence of control systems and the linearization of nonlinear systems," SIAM J. Control, Vol. 11, pp. 670-676, 1973.
8. R. W. Brockett, "Feedback invariants for nonlinear systems," IFAC Congress, Helsinki, 1978.
9. B. Jakubczyk and W. Respondek, "On linearization of control systems," Bull. Acad. Polon. Sci., Ser. Sci. Math. Astronom. Phys., to appear.
10. R. Su, "On the linear equivalents of nonlinear systems," submitted to Systems and Control Letters.

NONLINEAR FAULT ANALYSIS

11/2

Data base for symbolic network analysis

C.-C. Wu, M.S., and Prof. R. Saeks, M.S., Ph.D., Fel. I.E.E.E., P.E.

Indexing terms: Linear networks, Transfer functions

Abstract: A data base for generating the symbolic transfer functions for a linear electronic circuit is formulated and an appropriate retrieval theorem derived. The size of the required data base is $O(n^2)$ independently of the number of simultaneously varying parameters, where n is the total number of component output terminals, and the cost of retrieval is $O(p^3)$ multiplications where p is the actual number of circuit parameters which vary simultaneously in a given analysis. As such, both storage and computational requirements are minimised.

List of symbols

Matrix	Type	Dimension	Index
a	= composite component input vector	$m \times 1$	—
b	= composite component output vector	$n \times 1$	—
u	= composite system input vector	$v \times 1$	—
y	= composite system output vector	$q \times 1$	—
L_{11}	= connection matrix	$m \times n$	—
L_{21}	= connection matrix	$q \times n$	—
L_{21}^q	= q th row of L_{21}	$1 \times n$	$q = 1, 2, \dots, q$
L_{12}	= connection matrix	$m \times v$	—
L_{12}^v	= v th column of L_{12}	$m \times 1$	$v = 1, 2, \dots, v$
L_{22}	= connection matrix	$q \times v$	—
L_{22}^{qv}	= q - v entry in L_{22}	1×1	$q = 1, 2, \dots, q; v = 1, 2, \dots, v$
S	= composite system transfer function matrix	$q \times v$	—
S^{qv}	= q - v entry in S	1×1	$q = 1, 2, \dots, q; v = 1, 2, \dots, v$
Z	= composite component transfer-function matrix	$n \times m$	—
Z_0	= nominal composite component transfer-function matrix	$n \times m$	—
Z_1	= composite component transfer function perturbation matrix	$n \times m$	—
c^k	= column vector characterising perturbation of k th parameter	$n \times 1$	$k = 1, 2, \dots, k$
C	= array of the c^k vectors for the parameters which actually vary (row [c^k])	$n \times p$	—
r^k	= row vector characterising perturbation of k th parameter	$1 \times m$	$k = 1, 2, \dots, k$
R	= array of r^k vectors for the parameters which actually vary (col [r^k])	$p \times m$	—
δ^k	= k th variable parameter	1×1	$k = 1, 2, \dots, k$
Δ	= array of δ^k s for parameters which actually vary (diag [δ^k])	$p \times p$	—

1 Introduction

Historically, symbolic network analysis has been motivated by the problems of circuit design and, as such, the emphasis has been placed on quickly and efficiently obtaining a symbolic transfer function from a given set of circuit specifications [2, 3]. In an operational or maintenance environment, however, one is typically given a prescribed nominal circuit and desires to determine the effect of various (possibly large) perturbations thereon. This is the case in a power system where one is given a fixed network and desires to determine the effect of proposed modifications thereto. Alternatively, in the problem of analogue circuit fault diagnosis, one desires to simulate the effect of a number of alternative failures, to compare the simulated data with the observed failure data [4].

In such an operational or maintenance environment, numerous perturbations of the nominal circuit are studied and, as such, significant computational efficiencies can be obtained if one first generates a data base in terms of the nominal circuit parameters and then extracts the appropriate symbolic transfer function from the data base each time a different symbolic

transfer function is required. Of course, the benefit to be achieved via such an approach is dependent on the size of the data base and the ease with which a symbolic transfer function may be retrieved therefrom.

The obvious manner in which to generate such a data base is to simply precompute the coefficients of all required symbolic transfer functions and store them in the data base. Retrieval from such a data base is, of course, immediate, but the data base may become overly large. Indeed, the number of transfer functions which must be stored is $O(k^p)$ where k is the total number of potentially variable circuit parameters and p is the maximum number of circuit parameters which may vary simultaneously. An alternative approach is to store the nominal transfer function information and then use Householder's formula [1] to compute the required symbolic transfer functions. In such a data base we need only store $O(n^2)$ transfer functions, where n is the total number of component output terminals, but retrieval requires $O(n^3 + p^3)$ multiplications, where p is the actual number of circuit parameters which vary simultaneously. Since, in practice, $n \gg p$, the retrieval process requires approximately $O(n^3)$ multiplications and is dominated by the large dimensional matrix multiplication required by Householder's formula rather than the low-dimensional inverse.

In the present paper, we will formulate an alternative data base for the symbolic transfer functions which also requires

Paper 1525G, first received 21st November 1980 and in final form 18th June 1981

The authors are with the Departments of Electrical Engineering, Texas Tech University, Box 4439, Lubbock TE 79409, USA

multiplications. Since p is typically small, this is tantamount to immediate retrieval.

In the rest of this introduction, we will review the properties of the component connection model for a large-scale circuit or system [1, 6, 7] which serves as the starting point for our theory. The data base and retrieval formulas for the case where $p \leq 2$ are formulated in Section 2 and the general retrieval formula is derived in Section 3. Section 4 is devoted to the problem of retrieving sensitivity formulas from the data base and Section 5 deals with the problem of updating the data base when the nominal circuit parameters are changed. Finally, Section 6 is devoted to several examples illustrating the theory.

The component connection model is an algebraic model for an interconnected dynamical system which subsumes the classical topological models but is more readily manipulated both analytically and computationally [1, 6, 7]. The motivation and justification of the model are discussed in detail in Reference 1 and will not be repeated here. The component connection model takes the form of the set of simultaneous equations:

$$b = Z(j\omega)a \quad (1)$$

$$a = L_{11}b + L_{12}u \quad (2)$$

and

$$y = L_{21}b + L_{22}u \quad (3)$$

Here, $Z(=Z(j\omega))$ is a frequency-dependent matrix characterising the decoupled system components with composite component input and output vectors a and b , respectively. On the other hand the L_{ij} ; $i, j = 1, 2$; matrices are frequency-independent connection matrices characterising the coupling between the composite component vectors a and b and the composite system input and output vectors u and y , respectively.

A little algebra with the component connection equations will readily reveal that

$$S = L_{22} + L_{21}(1 - ZL_{11})^{-1}ZL_{12} \quad (4)$$

where $S(=S(j\omega))$ is the composite system transfer function matrix [1] characterising the external behaviour of the system via

$$y = S(j\omega)u \quad (5)$$

Often, rather than working with the entire S matrix, we find it convenient to work with its individual entries; s^{qv} , $q = 1, 2, \dots, q$ and $v = 1, 2, \dots, v$; which are related to the component connection model via

$$s^{qv} = L_{22}^{qv} + L_{21}^q(1 - ZL_{11})^{-1}ZL_{12}^v \quad (6)$$

Here L_{22}^{qv} is the q - v entry in L_{22} ; $q = 1, 2, \dots, q$ and $v = 1, 2, \dots, v$; L_{21}^q is the q th row of L_{21} ; $q = 1, 2, \dots, q$; and L_{12}^v is the v th column of L_{12} ; $v = 1, 2, \dots, v$.

Finally, since we are interested in analysing the effects of perturbing one or more components from their nominal values, we decompose Z into nominal and perturbation terms in the form

$$Z = Z_0 + Z_1 \quad (7)$$

where

$$Z_1 = \sum_{k=1}^p c^k \delta^k r^k \quad (8)$$

Here, $c^k(=c^k(j\omega))$ is a column vector, $r^k(=r^k(j\omega))$ is a row vector, and δ^k is the scalar perturbation for the k th potentially variable component parameter. In a typical application one is given c^k , r^k , and δ^k ; $k = 1, 2, \dots, k$; characterising k poten-

tially variable component parameters though at most p such parameters vary in any given analysis; $p \leq p \leq k$. Indeed, $p \leq k$ in most applications. Finally, we note that Z_1 can be expressed more concisely in the form

$$Z = C\Delta R \quad (9)$$

where

$$C = [c^1; c^2; \dots; c^p] \quad (10)$$

$$R = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^p \end{bmatrix} \quad (11)$$

and

$$\Delta = \begin{bmatrix} \delta^1 & & \\ & \delta^2 & \\ & & \ddots \\ & & & \delta^p \end{bmatrix} \quad (12)$$

The above-described notation, formulated for the component connection model, is summarised in the list of symbols:

2 Data base

Our data base is composed of the following family of (frequency dependent) scalar transfer functions:

$$s_0^{qv} = L_{22}^{qv} + L_{21}^q(1 - Z_0 L_{11})^{-1}Z_0 L_{12}^v \quad (13)$$

$$q = 1, 2, \dots, q; v = 1, 2, \dots, v$$

$$b^q = L_{21}^q(1 - Z_0 L_{11})^{-1}c^j \quad (14)$$

$$q = 1, 2, \dots, q; j = 1, 2, \dots, k$$

$$c^{kv} = r^k[1 + L_{11}(1 - Z_0 L_{11})^{-1}Z_0]L_{12}^v \quad (15)$$

$$k = 1, 2, \dots, k; v = 1, 2, \dots, v$$

and

$$e^{kj} = r^k L_{11}(1 - Z_0 L_{11})^{-1}c^j \quad (16)$$

$$k, j = 1, 2, \dots, k$$

Here, q and v denote the number of external system inputs and outputs which are typically few in number. As such, the e^{kj} array, composed of k^2 entries, dominates the data base. Also note that all the entries in the data base are formulated in terms of the nominal component values and, as such, the data base may be generated off-line without *a priori* knowledge of the perturbations to be analysed. Finally, the entire data base may be generated with the aid of only a single n by n (sparse) matrix inverse.

Now, if we assume that only a single parameter is perturbed, i.e.

$$Z_1 = c^k \delta^k r^k \quad (17)$$

for some fixed $k = 1, 2, \dots, k$, to retrieve s^{qv} from the data base we must evaluate

$$s^{qv} = L_{22}^{qv} + L_{21}^q(1 - [Z_0 + c^k \delta^k r^k]L_{11})^{-1} [Z_0 + c^k \delta^k r^k]L_{12}^v \quad (18)$$

in terms of the elements of our data base and the variable parameter δ^k . To this end, we invoke Householder's formula [1]

$$(W + XY)^{-1} = W^{-1} - W^{-1}X(1 + YW^{-1}X)^{-1}YW^{-1} \quad (19)$$

with

$$W = (1 - Z_0 L_{11}) \quad X = -c^k \delta^k \quad Y = r^k L_{11}$$

obtaining

$$\begin{aligned} (1 - [Z_0 + c^k \delta^k r^k] L_{11})^{-1} &= [(1 - Z_0 L_{11}) + (-c^k \delta^k) (r^k L_{11})]^{-1} \\ &= (1 - Z_0 L_{11})^{-1} + (1 - Z_0 L_{11})^{-1} c^k \delta^k (1 - r^k L_{11} (1 - Z_0 L_{11})^{-1} c^k \delta^k)^{-1} r^k L_{11} (1 - Z_0 L_{11})^{-1} \\ &= (1 - Z_0 L_{11})^{-1} + \frac{(1 - Z_0 L_{11})^{-1} c^k \delta^k r^k L_{11} (1 - Z_0 L_{11})^{-1}}{1 - \delta^k e^{kk}} \end{aligned} \quad (20)$$

Now, on substitution of eqn. 20 into eqn. 18, we obtain

$$\begin{aligned} s^{qv} &= L_{22}^{qv} + L_{21}^q (1 - [Z_0 + c^k \delta^k r^k] L_{11})^{-1} [Z_0 + c^k \delta^k r^k] L_{12}^v \\ &= L_{22}^{qv} + L_{21}^q (1 - Z_0 L_{11})^{-1} [Z_0 + c^k \delta^k r^k] L_{12}^v \\ &\quad + \frac{L_{21}^q (1 - Z_0 L_{11})^{-1} c^k \delta^k r^k L_{11} (1 - Z_0 L_{11})^{-1} [Z_0 + c^k \delta^k r^k] L_{12}^v}{1 - \delta^k e^{kk}} \\ &= s_0^{qv} + \delta^k b^{qk} r^k L_{12}^v + \frac{\delta^k b^{qk} r^k L_{11} (1 - Z_0 L_{11})^{-1} Z_0 L_{12}^v + (\delta^k)^2 b^{qk} e^{kk} r^k L_{12}^v}{1 - \delta^k e^{kk}} \\ &= s_0^{qv} + \frac{\delta^k b^{qk} d^{kv} + (\delta^k)^2 [-b^{qk} e^{kk} r^k L_{12}^v + b^{qk} e^{kk} r^k L_{12}^v]}{1 - \delta^k e^{kk}} = s_0^{qv} + \frac{\delta^k b^{qk} d^{kv}}{1 - \delta^k e^{kk}} \end{aligned} \quad (21)$$

which is the desired symbolic transfer function.

If we assume that two parameters are perturbed, i.e.

$$Z_1 = c^k \delta^k r^k + c^j \delta^j r^j \quad (22)$$

a similar formula can be obtained, wherein Householder's formula is applied twice. Since this formula is subsumed by the general retrieval formula derived in the following Section, we simply state the result without proof. In particular

$$\begin{aligned} s^{qv} &= L_{22}^{qv} + L_{21}^q (1 - [Z_0 + c^k \delta^k r^k + c^j \delta^j r^j] L_{11})^{-1} [Z_0 + c^k \delta^k r^k + c^j \delta^j r^j] L_{12}^v \\ &= s_0^{qv} + \frac{\delta^k b^{qk} d^{kv} + \delta^j b^{qj} d^{jv} + \delta^k \delta^j (-e^{kk} b^{qj} d^{jv} - e^{jj} b^{qk} d^{kv} + e^{kj} b^{qj} d^{kv} + e^{jk} b^{qk} d^{jv})}{1 - \delta^k e^{kk} - \delta^j e^{jj} + \delta^k \delta^j (e^{kk} e^{jj} - e^{kj} e^{jk})} \end{aligned} \quad (23)$$

3 Retrieval theorem

As is apparent from eqn. 23, our retrieval formulas are quite complex, even for the case $p = 2$ and, as such, a more compact notation is required if they are to be tractable. To this end, we assume that δ^k ; $k = 1, 2, \dots, p$ denote the potentially variable parameters, and that

$$Z_1 = c^k \delta^k r^k = C \Delta R \quad (24)$$

Of course, the same expression applies to any set of p potentially variable parameters, given an appropriate change of the index set. To obtain the required symbolic transfer function for

$$S = L_{22} + L_{21} (1 - [Z_0 + Z_1] L_{11})^{-1} [Z_0 + Z_1] L_{12} \quad (25)$$

with the above specified Z_1 , we now define the following matrices made up of elements from our data base:

$$S_0 = \begin{bmatrix} s_0^{11} & s_0^{12} & \dots & s_0^{1v} \\ s_0^{21} & s_0^{22} & \dots & s_0^{2v} \\ \vdots & \vdots & \ddots & \vdots \\ s_0^{p1} & s_0^{p2} & \dots & s_0^{pv} \end{bmatrix} \quad (26)$$

$$B = \begin{bmatrix} b^{11} & b^{12} & \dots & b^{1p} \\ b^{21} & b^{22} & \dots & b^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b^{p1} & b^{p2} & \dots & b^{pp} \end{bmatrix} \quad (27)$$

$$D = \begin{bmatrix} d^{11} & d^{12} & \dots & d^{1v} \\ d^{21} & d^{22} & \dots & d^{2v} \\ \vdots & \vdots & \ddots & \vdots \\ d^{p1} & d^{p2} & \dots & d^{pv} \end{bmatrix} \quad (28)$$

$$E = \begin{bmatrix} e^{11} & e^{12} & \dots & e^{1p} \\ e^{21} & e^{22} & \dots & e^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ e^{p1} & e^{p2} & \dots & e^{pp} \end{bmatrix} \quad (29)$$

and Δ is as defined by eqn. 12.

Theorem

Using the above notation,

$$\begin{aligned} S &= L_{22} + L_{21} (1 - [Z_0 + Z_1] L_{11})^{-1} [Z_0 + Z_1] L_{12} \\ &= S_0 + B (1 - \Delta E)^{-1} \Delta D \end{aligned}$$

Proof: First, we observe that

$$S_0 = L_{22} + L_{21}(1 - Z_0 L_{11})^{-1} Z_0 L_{12} \quad (30)$$

is just the nominal system transfer function matrix, and

$$B = L_{21}(1 - Z_0 L_{11})^{-1} C \quad (31)$$

and

$$\begin{aligned} D &= R[1 + L_{11}(1 - Z_0 L_{11})^{-1} Z_0] L_{12} \\ &= R(1 - L_{11} Z_0)^{-1} \end{aligned} \quad (32)$$

via Householder's formula. Finally,

$$E = R L_{11}(1 - Z_0 L_{11})^{-1} C$$

where R and C are as defined by eqns. 10 and 11. As such,

$$\begin{aligned} (1 - \Delta E)^{-1} &= (1 - \Delta R L_{11}(1 - Z_0 L_{11})^{-1} C)^{-1} \\ &= [1 + \Delta R L_{11}(1 - (1 - Z_0 L_{11})^{-1} C \Delta R L_{11})^{-1} \times \\ &\quad (1 - Z_0 L_{11})^{-1} C] \\ &= [1 + \Delta R L_{11}(1 - Z_0 L_{11} - Z_1 L_{11})^{-1} C] \\ &= [1 + \Delta R L_{11}(1 - Z L_{11})^{-1} C] \end{aligned} \quad (33)$$

where we have invoked Householder's formula with $Z = 1$, $X = \Delta R L_{11}$, and $Y = (1 - Z_0 L_{11})^{-1} C$; and eqn. 9. As such,

$$\begin{aligned} S_0 + B(1 - \Delta E)^{-1} \Delta D &= S_0 + L_{21}(1 - Z_0 L_{11})^{-1} \times \\ &\quad C[1 + \Delta R L_{11}(1 - Z L_{11})^{-1} C] \times \\ &\quad \Delta R(1 - L_{11} Z_0)^{-1} L_{12} \\ &= S_0 + L_{21}(1 - Z_0 L_{11})^{-1} [Z_1 + Z_1 L_{11}(1 - Z L_{11})^{-1} Z_1] \times \\ &\quad (1 - L_{11} Z_0)^{-1} L_{12} \\ &= S_0 + L_{21}(1 - Z_0 L_{11})^{-1} \times \\ &\quad \{[(1 - Z L_{11}) + Z_1 L_{11}](1 - Z L_{11})^{-1} Z_1(1 - L_{11} Z_0)^{-1} L_{12}\} \\ &= S_0 + L_{21}(1 - Z_0 L_{11})^{-1} (1 - Z_0 L_{11}) \times \\ &\quad (1 - Z L_{11})^{-1} Z_1(1 - L_{11} Z_0)^{-1} L_{12} \\ &= S_0 + L_{21}(1 - Z L_{11})^{-1} Z_1(1 - L_{11} Z_0)^{-1} L_{12} \\ &= L_{22} + L_{21}(1 - Z_0 L_{11})^{-1} Z_0 L_{12} \\ &\quad + L_{21}(1 - Z L_{11})^{-1} Z_1(1 - L_{11} Z_0)^{-1} L_{12} \\ &= L_{22} + L_{21} Z_0(1 - L_{11} Z_0)^{-1} L_{12} \\ &\quad + L_{21}(1 - Z L_{11})^{-1} Z_1(1 - L_{11} Z_0)^{-1} L_{12} \\ &= L_{22} + L_{21} [Z_0 + (1 - Z L_{11})^{-1} Z_1] (1 - L_{11} Z_0)^{-1} L_{12} \\ &= L_{22} + L_{21}(1 - Z L_{11})^{-1} [(1 - Z L_{11}) Z_0 + Z_1] \times \\ &\quad (1 - L_{11} Z_0)^{-1} L_{12} \\ &= L_{22} + L_{21}(1 - Z L_{11})^{-1} [Z - Z L_{11} Z_0] \times \\ &\quad (1 - L_{11} Z_0)^{-1} L_{12} \\ &= L_{22} + L_{21}(1 - Z L_{11})^{-1} Z [1 - L_{11} Z_0] \times \\ &\quad (1 - L_{11} Z_0)^{-1} L_{12} \\ &= L_{22} + L_{21}(1 - Z L_{11})^{-1} Z L_{12} = S \end{aligned}$$

as required.

4 Sensitivity formulas

If one is working directly with the component connection model, it is well known [4] that the sensitivity of S with

respect to a parameter δ^i can be computed via the formula

$$\left[\frac{dS}{d\delta^i} \right] = L_{21}(1 - Z L_{11})^{-1} \left[\frac{dZ}{d\delta^i} \right] [1 + L_{11}(1 - Z L_{11})^{-1} Z] L_{12}$$

and hence it is appropriate to ask whether or not such a sensitivity matrix can also be computed from our data base. Since the expression

$$S = S_0 + B(1 - \Delta E)^{-1} \Delta D \quad (35)$$

is formally identical to eqn. 4, if $1 \leq i \leq p$, we may write

$$\left[\frac{dS}{d\delta^i} \right] = B(1 - \Delta E)^{-1} M_i [1 + E(1 - \Delta E)^{-1} \Delta] D \quad (36)$$

where

$$M_i = \frac{d\Delta}{d\delta^i} \begin{bmatrix} 0 & & & & \\ & 0 & & & \\ & & 0 & & \\ & & & \ddots & \\ & & & & 1 \\ & & & & & 0 \end{bmatrix} \quad (37)$$

with the one appearing in the i th diagonal entry. Clearly, the expression can be computed directly from the data base with the same level of computational effort as required for the retrieval formula.

In the case where δ^i is not included in the given set of parameters which deviate from nominal, $i > p$ in our notation, we must first augment the B , E , D , and Δ matrices to include δ^i and then apply eqn. 36 to the augmented system. To this end, we let

$$B^i = \begin{bmatrix} b^{11} & b^{12} & \dots & b^{1p} & b^{1i} \\ b^{21} & b^{22} & \dots & b^{2p} & b^{2i} \\ \vdots & \vdots & & \vdots & \vdots \\ b^{q1} & b^{q2} & & b^{qp} & b^{qi} \end{bmatrix} \quad (38)$$

$$D^i = \begin{bmatrix} d^{11} & d^{12} & \dots & d^{1p} \\ d^{21} & d^{22} & \dots & d^{2p} \\ \vdots & \vdots & & \vdots \\ d^{p1} & d^{p2} & \dots & d^{pp} \\ d^{i1} & d^{i2} & \dots & d^{ip} \end{bmatrix} \quad (39)$$

$$E^i = \begin{bmatrix} e^{11} & e^{12} & \dots & e^{1p} & e^{1i} \\ e^{21} & e^{22} & \dots & e^{2p} & e^{2i} \\ \vdots & \vdots & & \vdots & \vdots \\ e^{p1} & e^{p2} & \dots & e^{pp} & e^{pi} \\ e^{i1} & e^{i2} & \dots & e^{ip} & e^{ii} \end{bmatrix} \quad (40)$$

and

$$\Delta^i = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \quad (41)$$

Then we obtain the retrieval formulas

$$S = S_0 + B^i(1 - \Delta^a E^i)^{-1} \Delta^a D^i \quad (42)$$

and

$$\left[\frac{dS}{d\delta^i} \right] = B^i(1 - \Delta^a E^i)^{-1} M_{p+1} [1 + E^i(1 - \Delta^a E^i)^{-1} \Delta^a] D^i \quad (43)$$

5 Updating the data base

In many applications, one uses a data base such as that described above, as a design tool to aid in simulating the effects of various proposed modifications to the system. When such a modification is finally implemented, it is then necessary to update the data base to reflect the new nominal parameter values

$$\dot{Z}_0 = Z_0 + \sum c^k \delta^k r^k = Z_0 + C \Delta R \quad (44)$$

With the aid of Householder's formula, we may compute

$$\begin{aligned} (1 - \tilde{Z}_0 L_{11}) &= [(1 - Z_0 L_{11}) - C \Delta R L_{11}]^{-1} \\ &= (1 - Z_0 L_{11})^{-1} + (1 - Z_0 L_{11})^{-1} \times \\ &\quad C [1 - \Delta R L_{11} (1 - Z_0 L_{11})^{-1} C]^{-1} \times \\ &\quad \Delta R L_{11} (1 - Z_0 L_{11})^{-1} \\ &= (1 - Z_0 L_{11})^{-1} + (1 - Z_0 L_{11})^{-1} \times \\ &\quad C (1 - \Delta E)^{-1} \Delta R L_{11} (1 - Z_0 L_{11})^{-1} \end{aligned} \quad (45)$$

which, on substitution into eqn. 16, yields

$$\tilde{e}^{kj} = e^{kj} + [e^{k1} e^{k2} \dots e^{kp}] (1 - \Delta E)^{-1} \Delta \begin{bmatrix} e^{1j} \\ e^{2j} \\ \vdots \\ e^{pj} \end{bmatrix} \quad (46)$$

Similarly,

$$\tilde{b}^{qj} = b^{qj} + [b^{q1} b^{q2} \dots b^{qp}] (1 - \Delta E)^{-1} \Delta \begin{bmatrix} e^{1j} \\ e^{2j} \\ \vdots \\ e^{pj} \end{bmatrix} \quad (47)$$

$$\tilde{d}^{kv} = d^{kv} + [e^{k1} e^{k2} \dots e^{kp}] (1 - \Delta E)^{-1} \Delta \begin{bmatrix} d^{1v} \\ d^{2v} \\ \vdots \\ d^{pv} \end{bmatrix} \quad (48)$$

and

$$\tilde{s}^{qv} = s^{qv} + [b^{q1} b^{q2} \dots b^{qp}] (1 - \Delta E)^{-1} \Delta \begin{bmatrix} d^{1v} \\ d^{2v} \\ \vdots \\ d^{pv} \end{bmatrix} \quad (49)$$

As such, the entries in our data base can be updated with a computational effort which is commensurate with that required by the retrieval formula.

6 Examples

Consider the simple RC op-amp circuit shown in Fig. 1. The component connection

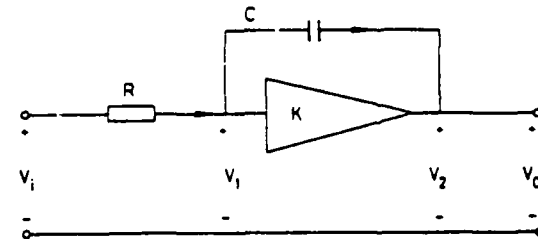


Fig. 1 RC op-amp circuit

model for this circuit takes the form

$$\begin{bmatrix} i_c \\ v_r \\ v_2 \end{bmatrix} = \begin{bmatrix} sC & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & K \end{bmatrix} \begin{bmatrix} v_c \\ i_r \\ v_1 \end{bmatrix} \quad (50)$$

$$\begin{bmatrix} v_c \\ i_r \\ v_1 \end{bmatrix} = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} i_c \\ v_r \\ v_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} v_i \quad (51)$$

$$\begin{bmatrix} v_o \\ - \\ - \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ & & v_r \\ & & v_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} v_i \quad (52)$$

Thus, if all components are taken to have nominal values of 1, we obtain

$$(1 - Z_0 L_{11}) = \begin{bmatrix} 1 & s & s \\ -1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad (53)$$

$$(1 - Z_0 L_{11})^{-1} = \begin{bmatrix} 1 & 0 & -s \\ 1 & 1 & -s \\ -1 & -1 & s+1 \end{bmatrix} \quad (54)$$

$$(1 - Z_0 L_{11})^{-1} Z_0 = \begin{bmatrix} s & 0 & -s \\ s & 1 & -s \\ -s & -1 & s+1 \end{bmatrix} \quad (55)$$

$$L_{11} (1 - Z_0 L_{11})^{-1} = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & -s \\ -1 & -1 & s \end{bmatrix} \quad (56)$$

and

$$1 + L_{11} (1 - Z_0 L_{11})^{-1} Z_0 = \begin{bmatrix} 1 & 0 & -s \\ s & 1 & -s \\ -s & -1 & s+1 \end{bmatrix}$$

Now, we may represent

FD-302 (Rev. 11-27-70)

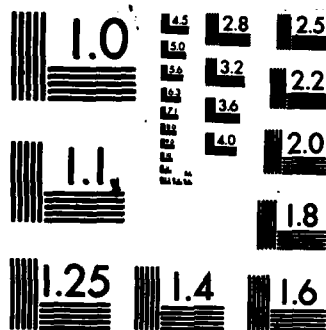
ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICES
ELECTRONICS PROGRAM VO. (U) TEXAS TECH UNIV LUBBOCK
INST FOR ELECTRONIC SCIENCE R SAEKS ET AL. DEC 82
N00014-76-C-1136 F/G 9/3

3/4

UNCLASSIFIED

F/G 9/3

Ni



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

and K via the matrices

$$c^1 \delta^1 r^1 = \begin{bmatrix} s \\ 0 \\ 0 \end{bmatrix} \delta^1 \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \quad (58)$$

$$c^2 \delta^2 r^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \delta^2 \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \quad (59)$$

and

$$c^3 \delta^3 r^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \delta^3 \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \quad (60)$$

Combining the appropriate c^h and r^j matrices with the above expressions as per eqns. 13 through 16, we obtain the data base

$$s_0 = 1 \quad (61)$$

$$b^1 = -s \quad b^2 = -1 \quad b^3 = s+1 \quad (62)$$

$$d^1 = 1 \quad d^2 = 0 \quad d^3 = 1 \quad (63)$$

and

$$\begin{aligned} e^{11} &= 0 & e^{12} &= 0 & e^{13} &= -1 \\ e^{21} &= s & e^{22} &= 0 & e^{23} &= -s \\ e^{31} &= -s & e^{32} &= -1 & e^{33} &= s \end{aligned} \quad (64)$$

where we have deleted the q and v indices, since we are dealing with a single-input/single-output system.

Now, if one desires to compute the symbolic transfer function with respect to perturbations in the op-amp gain, we have

Finally, if we desire to update our data base to reflect a new nominal value for the circuit parameters of $C = 1$, $R = 1$, and $K = 2$, we invoke eqns. 46 through 49 with $\delta^3 = 1$, yielding

$$\tilde{s}_0 = s_0 + \frac{b^3 \delta^3 d^3}{1 - \delta^3 e^{33}} \Big|_{\delta^3=1} = \frac{2}{1-s} \quad (67)$$

$$\tilde{e}^{11} = e^{11} + \frac{e^{13} \delta^3 e^{31}}{1 - \delta^3 e^{33}} = 0 + \frac{(-1)\delta^3(-s)}{1 - \delta^3(s)} \Big|_{\delta^3=1} = \frac{s}{1-s} \quad (68)$$

and similarly for the other elements of the data base.

As a more realistic example, consider the 2-stage feedback amplifier of Fig. 2, where, as before, we take all nominal component parameters to be one to simplify the illustration. The component connection model for this circuit takes the form:

$$\begin{bmatrix} v_o \\ i_{r_1} \\ i_{r_2} \\ v_{e_1} \\ v_{e_2} \\ v_o \\ i_{r_3} \\ i_{r_4} \\ v_{e_3} \\ v_{e_4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/s & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/s & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/s & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/s \end{bmatrix} \begin{bmatrix} v_{e_1} \\ v_{r_1} \\ v_{r_2} \\ i_{e_1} \\ i_{e_2} \\ v_{e_3} \\ v_{r_3} \\ v_{r_4} \\ i_{e_3} \\ i_{e_4} \end{bmatrix} \quad (69)$$

and

$$\begin{bmatrix} v_{e_1} \\ v_{r_1} \\ v_{r_2} \\ i_{e_1} \\ i_{e_2} \\ v_{e_3} \\ v_{r_3} \\ v_{r_4} \\ i_{e_3} \\ i_{e_4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_o \\ i_{r_1} \\ i_{r_2} \\ v_{e_1} \\ v_{e_2} \\ v_o \\ i_{r_3} \\ i_{r_4} \\ v_{e_3} \\ v_{e_4} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} v_i \quad (70)$$

$$S(s, \delta^3) = s_0 + \frac{b^3 \delta^3 d^3}{1 - \delta^3 e^{33}} = \frac{1 + \delta^3}{1 - \delta^3 s} \quad (67)$$

Recalling that δ^3 represents a perturbation from a nominal parameter value of $K_0 = 1$, our actual gain is $K = K_0 + \delta^3 = 1 + \delta^3$ which, on substitution into eqn. 65, yields

$$S(s, K) = \frac{K}{(1-K)s + 1} \quad (66)$$

which is the classical gain formula for such a circuit.

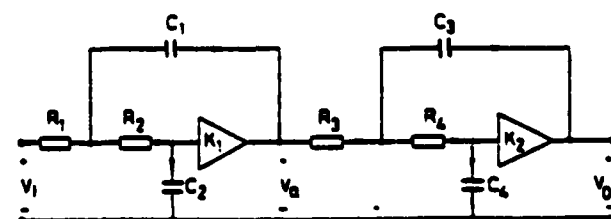


Fig. 2 2-stage feedback amplifier

$$v_0 = [0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0] \begin{bmatrix} v_a \\ i_{r_1} \\ i_{r_2} \\ v_{e_1} \\ v_{e_2} \\ v_0 \\ i_{r_3} \\ i_{r_4} \\ v_{e_3} \\ v_{e_4} \end{bmatrix}$$

$$+ [0] v_i \quad (71)$$

Now, after invoking the indicated operations, the entries in our data base for $q, j = 1, 2$, and 3 take the form:

$$s_0 = \frac{1}{(s+1)^4} \quad (72)$$

$$b^1 = \frac{s^2 + 3s + 1}{(s+1)^4} = \frac{1}{(s+1)^4} \quad b_3 = \frac{1}{(s+1)^3} \quad (73)$$

$$d^1 = \frac{1}{(s+1)^2} \quad d^2 = \frac{1}{(s+1)} \quad d^3 = \frac{s(s+3)}{(s+1)^2} \quad (74)$$

$$e^{11} = \frac{s}{(s+1)^2} \quad e^{12} = \frac{1}{(s+1)^2}$$

$$e^{13} = \frac{1}{(s+1)}$$

$$e^{21} = \frac{-s^3 - 5s^2 - 4s + 2}{(s+1)^2(s+2)} \quad e^{22} = \frac{-1}{(s+1)}$$

$$e^{23} = \frac{-2s-1}{(s+1)^2}$$

$$e^{31} = \frac{s^3 + 4s^2 + 2s - 2}{(s+1)^2(s+2)} \quad e^{32} = \frac{s}{(s+1)^2}$$

$$e^{33} = \frac{s}{(s+1)^2} \quad (75)$$

Note that the fact that $s_0 = 1/(s+1)^4$ implies that both stages of the amplifier induce a pair of poles at $s = -1$, as is well known for this circuit.

Now, let us consider the case where the gain of the first stage op amp is raised to 2. Since the nominal value is 1, we employ $\delta^1 = 1$ in the formula

$$S(s, \delta^1) = s_0 + \frac{b^1 \delta^1 d^1}{1 - \delta^3 e^{11}} \quad (76)$$

yielding

$$\begin{aligned} S(s, 1) &= \frac{1}{(s+1)^4} + \frac{\frac{s^2 + 3s + 1}{(s+1)^4}}{\left[1 - \frac{s}{(s+1)^2}\right]} \\ &= \frac{1}{(s+1)^4} \left[1 + \frac{s^2 + 3s + 1}{s^2 + s + 1}\right] \\ &= \frac{2}{s^2 + s + 1} \end{aligned}$$

indicating that the first stage of the amplifier now has a pair of complex conjugate poles at $-1/2 + \sqrt{3}/2$, as is expected for this circuit. Of course, the second stage is unchanged, retaining its poles at $s = -1$.

7 Conclusions

The preceding development has been motivated by operational and maintenance considerations rather than the design considerations. In such an environment, one typically deals with a fixed nominal system, but carries out repeated analyses thereon. As such, the cost of generating the required data base is secondary compared to the cost of storing the data base and retrieving information therefrom. In these respects, we believe that our data base is near optimal. Since the number of system inputs and outputs is typically small, our data base contains approximately k^2 elements (actually $k^2 + k(v+q) + vq$) where k is the total number of parameters which are potentially variable. This data base, however, contains sufficient information to permit one to retrieve symbolic transfer functions for any number $p \leq k$ of variable parameters. Indeed, the number of variable parameters in a symbolic transfer function is reflected only in the cost of retrieval which is of the order of p^3 multiplications (actually $p^3 + p^2 \cdot \frac{1}{2} p v(q \cdot \frac{1}{2} 1)$). Since p is typically small, say five or less, this is minimal.

8 References

- 1 DECARLO, R.A., and SAEKS, R.: 'Interconnected dynamical systems' (Marcel Dekker, New York) (to be published)
- 2 LIN, P.-M.: 'Symbolic network functions by a single path finding Algorithm'. Proceedings of the 7th Allerton conference on circuits and systems, University of Illinois, Oct. 1969, pp. 196-205
- 3 LIN, P.-M.: 'SNAP - a computer program for generating symbolic network functions'. Report TR-EE70-16, School of Electrical Engineering, Purdue University, Aug. 1970
- 4 PURL, N.N.: 'Symbolic fault diagnosis techniques'. In SAEKS, R., and LIBERTY, S.R. (Eds.) (Marcel Dekker, New York, 1977)
- 5 SAEKS, R., and CHAO, K.-S.: 'Continuations approach to large-scale sensitivity analysis', *IEEE J., Electron. Circ. & Syst.*, 1976, 1, (1), pp. 11-16
- 6 SINGH, S.P. and LIU, R.-W.: 'Existence of state equation representation of large-scale dynamical systems', *IEEE Trans.*, 1973, CT-20, pp. 339-346



Richard Saeks was born in Chicago, Illinois in 1941. He received the B.S. in 1964, the M.S. in 1965, and the Ph.D. in 1967 from Northwestern University, Evanston, IL, Colorado State University, Fort Collins, CO, and Cornell University, Ithaca, NY, respectively, all in electrical engineering.

He is presently Paul Whitfield Horn Professor of Electrical Engineering, Mathematics and Computer Science at Texas Tech University, Lubbock, TX, where he is involved in teaching and research in the areas of fault analysis, large-scale systems and mathematical system theory.

Dr. Saeks is a member of IEEE, AMS, SIAM, ASEE and Sigma Xi.



Chwan-chia Wu was born in Tainan, Taiwan in 1955. He received the B.S. in 1977, and the M.S. in 1979, from the National Chiao Tung University, Hsinchu, Taiwan, both in computer science. He is presently working on a Ph.D. in electrical engineering, at Texas Tech University, Lubbock, Texas, where his research deals with the fault diagnosis problem for ana-

11/11/17

Diagnosability of Nonlinear Circuits and Systems—Part II: Dynamical Systems

RICHARD SAEKS, FELLOW, IEEE, ALBERTO SANGIOVANNI-VINCENTELLI, MEMBER, IEEE, AND
V. VISVANATHAN, STUDENT MEMBER, IEEE

Abstract—A theory for the diagnosability of nonlinear dynamical systems, similar to the one in Part I [1] for memoryless systems, is developed. It is based on an input-output model of the system in a Hilbert space setting. A necessary and sufficient condition for the local diagnosability of the system, which is a rank test on a matrix, is derived. A simple sufficient condition is also derived. It is shown that, for locally diagnosable systems, there exist a finite number of test inputs that are sufficient to diagnose the system. Illustrative examples are presented.

Index Terms—Adjoint map, dynamical systems, Frechet derivative, Hilbert space, local diagnosability, measure.

I. INTRODUCTION

IN PART I [1], a theory for the diagnosability of nonlinear memoryless systems (dc circuits) was developed.¹ The theory consists of the following parts:

- 1) a necessary and sufficient condition for the local diagnosability of the system,
- 2) simplified tests for local diagnosability,
- 3) a theorem, establishing that, for a locally diagnosable system, there exists a finite number of test inputs that are sufficient to diagnose the system,
- 4) sufficient conditions for single fault diagnosability.

The main contribution of this paper is to develop a similar theory for the diagnosability of nonlinear *dynamical* circuits and systems. Similar to [1], we work with an input-output model of the system. Unlike the memoryless case, however, the inputs and outputs are in this case, functions of time and are hence properly considered to be elements of an infinite dimensional Hilbert space. However, since the parameter space remains finite dimensional, in spite of the finite dimensional setting, the theory yields a finite dimensional test matrix. Our objective here, therefore, is to indicate a mechanism whereby the existing diagnosability theory for linear systems [2] and nonlinear memoryless systems [1] can be extended to the

general nonlinear dynamical case. The sequence in which we present our results parallels that of [1].

II. PROBLEM FORMULATION

For a nonlinear dynamical system, given an input waveform, a set of initial conditions for the states of the system, and a set of parameter values, the output waveform is uniquely defined. For simplicity, we assume that the initial states are fixed at zero (if they are unknown they can be subsumed into the parameter set), and that all measurements are taken in a fixed time interval $[0, 1]$. Specifically, we consider systems with p input terminals, q output terminals, and k parameters that can be described by the equation

$$y = f(u, \alpha) \quad (1)$$

where, $u \in U \subset \mathcal{P}[0, 1]^p$, the space of \mathbb{R}^p -valued piecewise continuous functions of time on the interval $[0, 1]$, $y \in Y = \mathcal{P}[0, 1]^q$, and $\alpha \in A \subset \mathbb{R}^k$; A is open. Note that with an appropriate inner product, Y is a Hilbert space over the field of real numbers [4]. Also, we assume that $f: U \times A \rightarrow Y$ is continuous in u and continuously (Frechet) differentiable [5] in α .

At this point we introduce a simple example (see Fig. 1), which we will carry along with us to illustrate the various definitions and results in this paper: Note that the circuit contains both linear and nonlinear resistors and inductors, and as such, there exists no systematic technique to test the diagnosability of this circuit. The voltage $u(\cdot)$ is the input, while the measured output is the current $y(\cdot)$. The various branch relations are

$$i_d = I_0(e^{\beta u_d} - 1)$$

$$i_G = G u_G$$

$$i_r = \Gamma \phi_r$$

$$i_\gamma = \gamma \phi_\gamma^2$$

The set of parameters is given by

$$\alpha = [I_0 \beta G \Gamma \gamma]^T$$

where " T " denotes the transpose. The input-output model is given by the equation

$$y(t) = I_0(e^{\beta u(t)} - 1) + G u(t)$$

Manuscript received January 7, 1981; revised June 7, 1981. This work was supported in part by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136, and the Joint Services Electronics Program at the University of California at Berkeley under AFOSR Contract F49620-79-C-0178.

R. Saeks is with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

A. Sangiovanni-Vincentelli and V. Visvanathan are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

¹ A brief review of the analog fault diagnosis literature is available in [1].

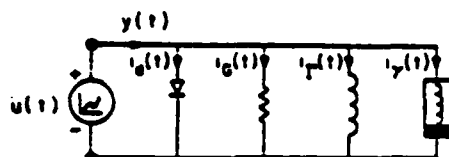


Fig. 1. The illustrative example.

$$+ \Gamma \int_0^1 u(\tau) d\tau + \gamma \left(\int_0^1 u(\tau) d\tau \right)^2. \quad (2)$$

We now present some definitions that are also available in Part I [1], but are relevant to the results presented here.

Definition 1: The parameter point $\alpha^0 \in A$ is said to be *locally diagnosable* if there exists an open neighborhood B of α^0 , such that $\forall \alpha \in B, \alpha \neq \alpha^0, \exists u \in U$, such that

$$f(u, \alpha^0) \neq f(u, \alpha). \quad \square$$

Definition 2: The system (1) is said to be *locally diagnosable* if almost all (i.e., all, except possibly those in some closed subset of A with zero Lebesgue measure) $\alpha \in A$ are locally diagnosable. \square

Definition 3: Let $M(\alpha)$ be a matrix whose elements are continuous functions of α everywhere in A . A parameter point

$$\begin{bmatrix} \int_0^1 (e^{\beta u(t)} - 1)^2 dt & \int_0^1 (e^{\beta u(t)} - 1)(I_0 u(t) e^{\beta u(t)}) dt \\ \int_0^1 (e^{\beta u(t)} - 1)(I_0 u(t) e^{\beta u(t)}) dt & \int_0^1 (I_0 u(t) e^{\beta u(t)})^2 dt \end{bmatrix}$$

$\alpha^0 \in A$ is said to be a *regular point* of $M(\alpha)$, if there exists an open neighborhood of α^0 in which $M(\alpha)$ has constant rank. \square

To define our test matrix, which is similar to the one introduced in [1], we let $J(u, \alpha)$ denote the Frechet derivative of f with respect to α , evaluated at u and α . With $u \in U$ fixed, f maps A to Y , hence, $J(u, \alpha)$ is a linear transformation² from \mathbb{R}^k to Y . $J(u, \alpha)$ can be described by the following map:

$$J(u, \alpha): a \rightarrow \left[\frac{\partial f}{\partial \alpha}(u, \alpha) \right] a \quad \forall a \in \mathbb{R}^k$$

where $\frac{\partial f}{\partial \alpha}(u, \alpha)$ is the matrix of partial derivatives of f with respect to the components of α , computed in the usual way and evaluated at (u, α) . For our example

$$\frac{\partial f}{\partial \alpha}(u, \alpha) = \begin{bmatrix} e^{\beta u(t)} - 1 & I_0 u(t) e^{\beta u(t)} \\ \int_0^1 u(\tau) d\tau & \left(\int_0^1 u(\tau) d\tau \right)^2 \end{bmatrix} \quad (3)$$

Note that, in (3), when $\frac{\partial f}{\partial \alpha}(u, \alpha)$ is multiplied on the right by an element of \mathbb{R}^k , i.e., the parameter space, the result is a scalar function of time, i.e., an element of Y . However, (3) is not the matrix representation of the linear operator $J(u, \alpha)$.

Let $J^*(u, \alpha)$ denote the adjoint map³ of $J(u, \alpha)$ [4]. Due

to the self duality of Y and \mathbb{R}^k , $J^*(u, \alpha)$ is a linear map from Y to \mathbb{R}^k [4]. To derive the adjoint map, we first define the inner product in the Hilbert space Y by

$$(x|y) = \int_0^1 x(t) \tau y(t) dt \quad \forall x, y \in Y$$

and in \mathbb{R}^k , in the usual way

$$(a|b) = a^T b \quad \forall a, b \in \mathbb{R}^k.$$

It now follows from the definition of the adjoint map [4] that

$$J^*(u, \alpha): y \rightarrow \int_0^1 \frac{\partial f}{\partial \alpha}(u(t), \alpha) \tau y(t) dt \quad \forall y \in Y.$$

Thus, $J^*(u, \alpha)J(u, \alpha)$ is a linear map from \mathbb{R}^k to \mathbb{R}^k , and can be directly identified with its matrix representation by the following equation:

$$J^*(u, \alpha)J(u, \alpha) = \int_0^1 \frac{\partial f}{\partial \alpha}(u(t), \alpha) \tau \frac{\partial f}{\partial \alpha}(u(t), \alpha) dt. \quad (4)$$

Note that $J^*(u, \alpha)J(u, \alpha)$ is a symmetric, positive semidefinite matrix. Observe that for our example, the top left 2×2 submatrix of $J^*(u, \alpha)J(u, \alpha)$ is given by

$$\begin{bmatrix} \int_0^1 (e^{\beta u(t)} - 1)(I_0 u(t) e^{\beta u(t)}) dt & \int_0^1 (I_0 u(t) e^{\beta u(t)})^2 dt \\ \int_0^1 (e^{\beta u(t)} - 1)(I_0 u(t) e^{\beta u(t)}) dt & \int_0^1 (I_0 u(t) e^{\beta u(t)})^2 dt \end{bmatrix}$$

The remaining elements of the matrix are similarly defined. Note that $J^*(u, \alpha)J(u, \alpha)$ is a matrix-valued function of u and α . To derive a test matrix that depends only on α , we integrate $J^*(u, \alpha)J(u, \alpha)$ over all possible inputs. To this end, let ω denote a positive measure defined on the Borel sets of U [6], such that $\omega(V) > 0$, for every nonnull open set V in U . Then we define the test matrix

$$R(\alpha) \triangleq \int_U J^*(u, \alpha)J(u, \alpha) d\omega(u). \quad (5)$$

III. CONDITIONS FOR LOCAL DIAGNOSABILITY

In this section, we first present a theorem that gives a necessary and sufficient condition for the local diagnosability of a parameter point. We then extend this theorem to give a condition for the local diagnosability of the system.

Theorem 1: Let $\omega(\cdot)$ be any admissible measure for which $R(\alpha)$ exists $\forall \alpha \in A$, and let α^0 be a regular point of $R(\alpha)$. Under these conditions, the parameter point α^0 is locally diagnosable if and only if $R(\alpha^0)$ is nonsingular.

Proof—If: By the integral form of the mean value theorem [7] we have

$$f(u, \alpha) - f(u, \alpha^0) = \int_0^1 \frac{\partial f}{\partial \alpha}(u, s\alpha^0 + (1-s)\alpha) ds [\alpha - \alpha^0] \quad \forall u \in U, \alpha \in A. \quad (6)$$

Suppose now that α^0 is not locally diagnosable. Then there exists an infinite sequence of vectors, $\alpha^i \rightarrow \alpha^0, i = 1, \dots, \infty$, such that

² Intuitively, the Frechet derivative is a tangent plane approximation to the function at the point at which it is evaluated.

³ A generalization of the concept of the transpose of a matrix.

$$f(u, \alpha^i) = f(u, \alpha^0) \quad \forall u \in U, \quad \forall i \in N.$$

Using (6) we have, for all u and i

$$\int_0^1 \frac{\partial f}{\partial \alpha} (u, s\alpha^0 + (1-s)\alpha^i) ds [\alpha^i] = 0 \quad (7)$$

where $\alpha^i = [\alpha^i - \alpha^0] / \|\alpha^i - \alpha^0\|$. Since α^i is normalized to lie on the unit sphere of \mathbb{R}^k , which is a compact set, α^i admits a convergent subsequence α^{i_j} , whose limit a also has unit norm. Using the convergent subsequence in (7), we have

$$\begin{aligned} \lim_{j \rightarrow \infty} \int_0^1 \frac{\partial f}{\partial \alpha} (u, s\alpha^0 + (1-s)\alpha^{i_j}) ds [\alpha^{i_j}] \\ = \int_0^1 \frac{\partial f}{\partial \alpha} (u, \alpha^0) ds [a] = \frac{\partial f}{\partial \alpha} (u, \alpha^0) a = 0. \end{aligned}$$

Since a has unit norm, it is nonzero, while

$$\begin{aligned} -a^T R(\alpha^0) a &= \int_U \int_0^1 \left[\frac{\partial f}{\partial \alpha} (u(t), \alpha^0) a \right]^T \\ &\quad \times \left[\frac{\partial f}{\partial \alpha} (u(t), \alpha^0) a \right] dt d\omega(u) = 0 \end{aligned}$$

implying that $R(\alpha^0)$ is nonsingular.

Only if: Conversely, if $R(\alpha^0)$ is singular it follows from the assumption that α^0 is a regular point, and Lemma 1 of [3]⁴ that there exists an open neighborhood V of α^0 and a continuous \mathbb{R}^k valued function $c(\alpha) \neq 0$ defined on V such that

$$\begin{aligned} 0 = c^T(\alpha) R(\alpha) c(\alpha) &= \int_U \int_0^1 \left[\frac{\partial f}{\partial \alpha} (u(t), \alpha) c(\alpha) \right]^T \\ &\quad \times \left[\frac{\partial f}{\partial \alpha} (u(t), \alpha) c(\alpha) \right] dt d\omega(u). \end{aligned}$$

Since $\omega(u) > 0$ and $\frac{\partial f}{\partial \alpha} (u, \alpha)$ is continuous in u , this implies that

$$\frac{\partial f}{\partial \alpha} (u, \alpha) c(\alpha) = 0 \quad \forall \alpha \in V, \quad \forall u \in U.$$

Finally, we define a curve $\alpha(s) \in V$ by the differential equation

$$\frac{\partial \alpha}{\partial s} = c(\alpha); \quad \alpha(0) = \alpha^0.$$

Substituting $\alpha(s)$ into $f(u, \alpha)$ and computing its derivative with respect to s via the chain rule [7], we obtain

$$\frac{\partial f}{\partial s} (u, \alpha(s)) = \frac{\partial f}{\partial \alpha} (u, \alpha(s)) \frac{\partial \alpha}{\partial s} = \frac{\partial f}{\partial \alpha} (u, \alpha(s)) c(\alpha) = 0$$

showing that $f(u, \alpha(s))$ is constant along a curve emanating from α^0 . Since $\alpha(s)$ is independent of u , this implies that α^0 is not locally diagnosable, thereby completing the proof. \square

For brevity, in the above theorem, we have used a proof that is similar to the proof of Theorem 1 of [3] and is shorter than the corresponding one in Part I [1]. Note that the proof uses

⁴ This fact is also established in the initial steps of the proof of Lemma 1 of [1].

the finite dimensionality of the parameter space, but does not require that U and Y be finite dimensional. Furthermore, the result is independent of the choice of measure ω .

Remark 1: In exactly the same way as has been done for linear systems [2] and nonlinear memoryless systems [1], we can establish that

$$\mu(\alpha^0) \triangleq k - \text{rank} [R(\alpha^0)]$$

is the measure of solvability of the parameter point α^0 , and that its generic value μ is the measure of testability of the system. \square

The rank test on $R(\alpha)$ becomes a test for the local diagnosability of the system, under exactly the same conditions as in the dc case. This is immediately obvious, since in both cases the analysis is restricted to the finite dimensional parameter space.

Theorem 2: Suppose that $f(u, \alpha)$ is analytic with respect to α . Then

- 1) almost all $\alpha \in A$ are regular points of $R(\alpha)$,
- 2) the system is locally diagnosable if and only if generic $\text{rank} R(\alpha) = k$. \square

IV. SIMPLIFIED TEST FOR LOCAL DIAGNOSABILITY AND EXISTENCE OF TEST INPUTS

In the preceding section, we have discussed the conditions under which the test for the local diagnosability of a system reduces to a rank test on the matrix $R(\alpha^0)$, evaluated at a randomly chosen point α^0 . Similar to the development in [1], our next step is to derive a simple sufficient condition for the local diagnosability of a parameter point.

Proposition 1: Suppose that there exists m inputs $u^i, i = 1, \dots, m \in U$ such that the matrix

$$\sum_{i=1}^m J^*(u^i, \alpha^0) J(u^i, \alpha^0)$$

is nonsingular. Then α^0 is locally diagnosable.

Proof: With $u^i, i = 1, \dots, m$ fixed, the function⁵

$$\text{col} [f(u^i, \alpha), i = 1, \dots, m]$$

is a map from A to Y^m . By the inverse function theorem [7], and the definition of local diagnosability, α^0 is locally diagnosable, if the linear map

$$\text{col} [J(u^i, \alpha^0), i = 1, \dots, m]$$

from \mathbb{R}^k to Y^m , is injective. This is true if and only if⁶ [7]

$$\text{Null space} [\text{col} (J(u^i, \alpha^0), i = 1, \dots, m)] = \{0\}$$

$$\Rightarrow \dim \text{range} [\text{row} (J^*(u^i, \alpha^0), i = 1, \dots, m)] = k$$

$$\Rightarrow \dim \text{range} [\text{row} (J^*(u^i, \alpha^0), i = 1, \dots, m) \cdot \text{col} (J(u^i, \alpha^0), i = 1, \dots, m)] = k$$

$$\Rightarrow \det \left[\sum_{i=1}^m J^*(u^i, \alpha^0) J(u^i, \alpha^0) \right] \neq 0. \quad \square$$

We now consider, once again, the illustrative example.

⁵ $\text{col}(A, B) \triangleq \begin{bmatrix} A \\ B \end{bmatrix}$, $\text{row}(A, B) \triangleq [A \ B]$.

Recall (3) and (4). As an example for the above proposition, we evaluate $J^*(u, \alpha)J(u, \alpha)$ for the input

$$u^0(t) = t \quad \forall t \in [0, 1]$$

and the parameter values

$$\alpha^0 = [1 \ 1 \ 1 \ 1 \ 1]^T.$$

Note that

$$\frac{\partial f}{\partial \alpha}(u^0, \alpha^0) = [e^t - 1; t e^t; t^2/2; t^3/2]$$

and

$$J^*(u^0, \alpha^0)J(u^0, \alpha^0) = \begin{bmatrix} 0.75796 & 1.09726 & 0.50000 & 0.19247 & 0.06613 \\ 1.09726 & 1.59726 & 0.33333 & 0.28172 & 0.09890 \\ 0.50000 & 0.33333 & 0.33333 & 0.12500 & 0.04167 \\ 0.19247 & 0.28172 & 0.12500 & 0.05000 & 0.01786 \\ 0.06613 & 0.09890 & 0.04167 & 0.01786 & 0.00694 \end{bmatrix}$$

Note that the above matrix is nonsingular. Hence, α^0 is locally diagnosable. Further, since the input-output map (2) is analytic with respect to α , the system is locally diagnosable.

Example 1: Consider the circuit of Fig. 2. Note that the circuit is "degenerate," since we are applying a voltage source across a capacitor. It is included here solely because it provides a simple but interesting illustration of Proposition 1. Because of the degeneracy of the circuit, and since we assumed zero initial conditions, we restrict our inputs to continuous function of time that satisfy the equation

$$u(0) = 0.$$

For the circuit of Fig. 2, the voltage $u(\cdot)$ is the input and the current $y(\cdot)$, the output. The various branch relations are

$$J^*(u^1, \alpha^0)J(u^1, \alpha^0) = \begin{bmatrix} 0.75796 & 1.09726 & 0.50000 & 0.50000 & 0.19247 \\ 1.09726 & 1.59726 & 0.33333 & 0.33333 & 0.28172 \\ 0.50000 & 0.33333 & 0.33333 & 0.33333 & 0.12500 \\ 0.50000 & 0.33333 & 0.33333 & 0.33333 & 0.12500 \\ 0.19247 & 0.28172 & 0.12500 & 0.12500 & 0.05000 \end{bmatrix}$$

$$I_d = I_0(e^{\beta u} - 1)$$

$$I_G = G u$$

$$q_C = \frac{C}{2} u^2$$

$$i_T = K \phi_T.$$

Hence, the input-output model is given by the equation

$$y(t) = I_0(e^{\beta u(t)} - 1) + G u(t) + C u(t) \frac{du}{dt}(t) + \Gamma \int_0^t u(\tau) d\tau. \quad (8)$$

$$J^*(u^2, \alpha^0)J(u^2, \alpha^0) = \begin{bmatrix} 0.43930 & 0.62862 & 0.29457 & 0.50023 & 0.08337 \\ 0.62862 & 0.90564 & 0.41755 & 0.71860 & 0.11977 \\ 0.29457 & 0.41755 & 0.20000 & 0.33333 & 0.05556 \\ 0.50023 & 0.71860 & 0.33333 & 0.57143 & 0.09524 \\ 0.08337 & 0.11977 & 0.05556 & 0.09524 & 0.01587 \end{bmatrix}$$

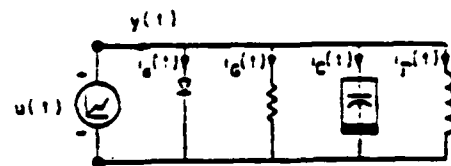


Fig. 2. Circuit for Example 1.

Note that

$$\alpha = [I_0 \ \beta \ G \ C \ \Gamma]^T.$$

We set

$$\alpha^0 = [1 \ 1 \ 1 \ 1 \ 1]^T$$

and note that

$$\frac{\partial f}{\partial \alpha}(u, \alpha^0) = \left[e^{u(t)} - 1; u(t)e^{u(t)}; u(t) \frac{du}{dt}(t); \int_0^t u(\tau) d\tau \right]. \quad (9)$$

To study the diagnosability of this system, we try the test input

$$u^1(t) = t \quad \forall t \in [0, 1].$$

Evaluating (9) for $u = u^1$, we have

$$\frac{\partial f}{\partial \alpha}(u^1, \alpha^0) = [e^t - 1; t e^t; t^2/2; t^3/2]$$

and

Note that the rank of the above matrix is 4 and in fact, the third and the fourth columns are linearly dependent. Equivalently, if the columns of $\frac{\partial f}{\partial \alpha}(u^1, \alpha^0)$ are considered to be elements of a vector space, the third and fourth columns are linearly dependent.

Consider now the possible test input

$$u^2(t) = t^2 \quad \forall t \in [0, 1]$$

$$\frac{\partial f}{\partial \alpha}(u^2, \alpha^0) = [e^{t^2} - 1; t^2 e^{t^2}; t^2; 2t^3; t^3/3]$$

and

Notice that, once again, the rank is 4 and in this case it is the fourth and fifth columns that are linearly dependent. However,

$$\sum_{i=1}^2 J^*(u^i, \alpha^0) J(u^i, \alpha^0)$$

is nonsingular, establishing that the system is locally diagnosable. Finally, it is easy to verify that for

$$u^3(t) = t^3 \quad \forall t \in [0, 1]$$

the matrix

$$J^*(u^3, \alpha^0) J(u^3, \alpha^0)$$

is nonsingular.

Once it has been verified that a system is locally diagnosable, it still remains to pick a set of test signals $\{u^i, i = 1, \dots, m\}$, and to solve the resulting set of equations

$$y^i = f(u^i, \alpha) \quad i = 1, \dots, m \quad (10)$$

for $\alpha \in A$. For this purpose we require that, for a representative parameter point α^0 , the linear map from \mathbb{R}^k to Y^m ,

$$H_m \triangleq \text{col}[J(u^i, \alpha^0), i = 1, \dots, m]$$

be injective, so that a Newton-Raphson type of algorithm will be assured to converge to α^0 , from a sufficiently good initial guess. More generally, if the property of injectiveness is generic in α , almost all faults can be diagnosed with this set of test inputs. Equivalently, since H_m can be identified with the time-varying matrix⁶

$$F_m \triangleq \begin{bmatrix} \frac{\partial f}{\partial \alpha_1}(u^1, \alpha^0) & \dots & \frac{\partial f}{\partial \alpha_k}(u^1, \alpha^0) \\ \vdots & & \vdots \\ \frac{\partial f}{\partial \alpha_1}(u^m, \alpha^0) & \dots & \frac{\partial f}{\partial \alpha_k}(u^m, \alpha^0) \end{bmatrix}$$

and the columns of F_m may be considered to be elements of Y^m , we require that these column vectors be linearly independent. For ease of exposition we will refer to this condition as F_m having column rank k .

If Proposition 1 is used to test for local diagnosability, then the inputs $\{u^i, i = 1, \dots, m\}$ that are used to establish local diagnosability, will also suffice as test inputs. More generally, we have the following theorem.

Theorem 3: Let ω be an admissible measure for which $R(\alpha)$ exists, $\forall \alpha \in A$, and suppose that $R(\alpha^0)$ is nonsingular at a regular point $\alpha^0 \in A$. Then \exists a sequence $u^i \in U, i = 1, \dots, m \leq k$, for which F_m has column rank k .

Proof: If $\frac{\partial f}{\partial \alpha_1}(u, \alpha^0) = 0$ (the zero element in Y), $\forall u \in U$, it follows from the definition of $R(\alpha^0)$ [see (4) and (5)], that the first row and the first column of $R(\alpha^0)$ is zero. This, however, contradicts the assumption that $R(\alpha^0)$ is nonsingular, and we may therefore assume that there exists a $u^1 \in U$ for which $\frac{\partial f}{\partial \alpha_1}(u^1, \alpha^0) \neq 0$. As such, there exists $u^1 \in U$ such that

F_1 has a column rank of at least 1. Using this fact as a starting point for an inductive hypothesis, we assume that there exists $u^i, i = 1, \dots, n$, such that the matrix F_n has column rank $j < k$, where $n \leq j$. We now desire to verify the existence of a vector $u^{n+1} \in U$ for which the corresponding matrix F_{n+1} has column rank greater than or equal to $j + 1$. To this end we let D be a nonsingular matrix of scalars which operate on the columns of F_n in such a way that the $(j + 1)$ st column through the k th column of $F_n D$ is zero. Since D is nonsingular, F_{n+1} will have column rank greater than, or equal to, $j + 1$ for some u^{n+1} if and only if $F_{n+1} D$ has column rank greater than, or equal to, $j + 1$. Because of the special form of $F_n D$, however, this will be the case, if and only if, the bottom row of $F_{n+1} D$ given by

$$\frac{\partial f}{\partial \alpha}(u^{n+1}, \alpha^0) D$$

is nonzero in columns $j + 1$ through k , for some u^{n+1} . If this is not the case, we may let d denote the $(j + 1)$ st column of D which is nonzero since D is nonsingular, in which case, we have

$$\frac{\partial f}{\partial \alpha}(u, \alpha^0) d = 0 \quad \forall u \in U.$$

This, however, implies that

$$d^T R(\alpha^0) d = \int_U \int_0^1 \left[\frac{\partial f}{\partial \alpha}(u(t), \alpha^0) d \right]^T \times \left[\frac{\partial f}{\partial \alpha}(u(t), \alpha^0) d \right] dt d\omega(u) = 0$$

which contradicts the assumption that $R(\alpha^0)$ is nonsingular. As such, there must exist u_{n+1} for which H_{n+1} has column rank greater than, or equal to $j + 1$. Repeating the argument inductively until an H_m with column rank k is obtained, now completes the proof of the theorem. Note that since $n \leq j$ at each step, $m \leq k$.

Remark 2: Recall from our examples, that unlike the linear dynamical [2] and the nonlinear memoryless [1] cases, a single-input, single-output nonlinear dynamical system can be tested with fewer test input signals than the number of parameters. \square

Note finally, that the results of [1] on single fault diagnosability can also be extended to the dynamical case. This is extremely straightforward, since most of the analysis is in the parameter space, which is finite dimensional.

V. CONCLUSIONS

Our purpose in this paper has been to indicate a mechanism whereby the existing diagnosability theory for linear systems [2] and memoryless nonlinear systems [1] can be extended to the general nonlinear dynamical case. To this end we have presented a necessary and sufficient condition for the local diagnosability of nonlinear dynamical systems described by the explicit input-output model

$$y = f(u, \alpha) \quad (11)$$

where u and y are elements of infinite dimensional Hilbert spaces, while α belongs to $A \subset \mathbb{R}^k$. On the basis of this con-

⁶ a_j is the j th entry in the vector α .

dition, we have shown that for a diagnosable system there exists a finite number of test inputs that are sufficient to diagnose the system.

However, in the case of circuits and systems of any reasonable size, the explicit relationship (11) is difficult to obtain, and usually only the implicit equation

$$g(\dot{x}(t), x(t), y(t), u(t), \alpha) = 0 \quad \forall t \geq 0 \quad (12)$$

is available [8]. In the above equation, $x(t)$ represents a set of internal variables and $\forall t \geq 0$, $x(t) \in \mathbb{R}^l$, $y(t) \in \mathbb{R}^q$, and $u(t) \in \mathbb{R}^p$. It follows immediately from the fact that

$$\dot{x}(t) = Dx(t)$$

where D is the differentiation operator, that (12) is equivalent to the operator relation

$$G(x, y, u, \alpha) = 0 \quad (13)$$

where, now, x , y , and u are elements of infinite dimensional Hilbert spaces. The fact that only the relation (13), that implicitly defines (11), is available, poses no theoretical problems.

It is obvious from our derivation in the previous sections, that we do not require that the *symbolic* form of the input-output model be known. We only require that given $(\hat{u}, \hat{\alpha}) \in U \times A$, $f(\hat{u}, \hat{\alpha})$, and $J(\hat{u}, \hat{\alpha})$, the Frechet derivative of f with respect to α evaluated at $(\hat{u}, \hat{\alpha})$, can be determined. A solution of (13) or equivalently (12) with $(u, \alpha) = (\hat{u}, \hat{\alpha})$ will give us $f(\hat{u}, \hat{\alpha})$, while an application of the implicit function theorem [7] on (13) gives us $J(\hat{u}, \hat{\alpha})$. In using the implicit function theorem however, we need to find the inverse of the Frechet derivative of G with respect to x and y . This is not a computationally easy task. A similar problem arises if we attempt to implement a diagnosis algorithm that solves (10) directly using a Newton-like technique. From a computational point of view therefore, we find that it is better to work with (12). In fact by studying a discrete-time version of (12), i.e., with the variable $\dot{x}(t)$ in (12) replaced by a numerical backward differentiation formula, as is done in a circuit simulator, it is possible to derive a sufficient condition for the diagnosability of a large class of dynamical systems, that is easily computed, and leads directly to an implementable diagnosis algorithm.

REFERENCES

- [1] V. Visvanathan and A. Sangiovanni-Vincentelli, "Diagnosability of nonlinear circuits and systems—Part I: The dc case," this issue, pp. 889–898.
- [2] N. Sen and R. Seaks, "Fault diagnosis for linear systems via multifrequency measurements," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 457–465, July 1979.
- [3] V. Visvanathan and A. Sangiovanni-Vincentelli, "Fault diagnosis of nonlinear memoryless systems," in *Proc. IEEE Int. Symp. Circuits Syst.*, Houston, TX, 1980, pp. 1082–1086.
- [4] A. E. Taylor and D. C. Lay, *Introduction to Functional Analysis*, 2nd ed. New York: Wiley, 1980.
- [5] J. T. Schwartz, *Nonlinear Functional Analysis*. New York: Gordon and Breach, 1969.
- [6] P. R. Halmos, *Measure Theory*. New York: Springer, 1974.
- [7] J. Dieudonné, *Foundations of Modern Analysis*. New York: Academic, 1960.
- [8] G. D. Hachtel, R. K. Brayton, and F. G. Gustavson, "The sparse tableau approach to network analysis and design," *IEEE Trans. Circuit Theory*, vol. CT-18, pp. 101–113, 1971.



Richard Seaks (S'59–M'65–SM'74–F'77) was born in Chicago, IL, in 1941. He received the B.S. degree in 1964, the M.S. degree in 1965, and the Ph.D. degree in 1967 from Northwestern University, Evanston, IL, Colorado State University, Fort Collins, and Cornell University, Ithaca, NY, respectively, all in electrical engineering.

He is presently Paul Whitfield Horn Professor of Electrical Engineering and Mathematics at Texas Tech University, Lubbock, where he is involved in teaching and research in the areas of fault analysis, circuit theory, and mathematical system theory.

Dr. Seaks is a member of the American Mathematical Society, the Society for Industrial and Applied Mathematics, and Sigma XI.

Alberto Sangiovanni-Vincentelli (M'74), for a biography, see this issue, page 898.

V. Visvanathan (S'78), for a photograph and biography, see this issue, page 898.

Analog Fault Diagnosis with Failure Bounds

CHWAN-CHIA WU, KAZUO NAKAJIMA, MEMBER, IEEE, CHIN-LONG WEY, AND
RICHARD SAEKS, FELLOW, IEEE

Abstract—A simulation-after-test algorithm for the analog fault diagnosis problem is proposed in which a bound on the maximum number of simultaneous failures is used to minimize the number of test points required. The resultant algorithm is applicable to both linear and nonlinear systems with multiple hard or soft faults and can be used to isolate failures up to an arbitrarily specified "replaceable chip or subsystem."

I. INTRODUCTION

CONCEPTUALLY, analog fault diagnosis algorithms can be subdivided into three classes [4]: simulation-before-test, simulation-after-test with a single test vector, and simulation-after-test with multiple test vectors. The former is commonly employed in digital testing and is characterized by minimal on-line computational requirements. Unfortunately, the high cost of analog circuit simulation coupled with the large number of potential fault modes which must be simulated in an analog circuit limits the applicability of simulation-before-test algorithms in an analog test environment. As an alternative to simulation-before-test, a number of researchers have proposed simulation-after-test algorithms, in which the internal system variables or component parameters are computed from the test data via a "nonlinear equation solver-like" algorithm. Indeed, in the case where sufficiently many test points are available only a single test vector is required and the fault diagnosis problem reduces to the solution of a linear equation [12], [15]. Except for the large number of test points required, this approach is ideally suited to the analog fault diagnosis problem and, as such, a considerable research effort has been directed towards the problem of reducing its test point requirements [4], [10], [13], [14], [16]. One such approach uses multiple test vectors to increase the number of equations obtained from a given set of test points [3], [7]. Unfortunately, this is achieved at the cost of greatly complicating the set of simultaneous equations which must be solved and, as such, the applicability of the approach is limited.

The purpose of the present paper is to describe an alternative simulation-after-test algorithm in which a bound on the maximum number of simultaneous failures is used to reduce the test point requirements while still retaining the computational simplicity inherent in a single test vector algorithm. Indeed, even though a given circuit may contain

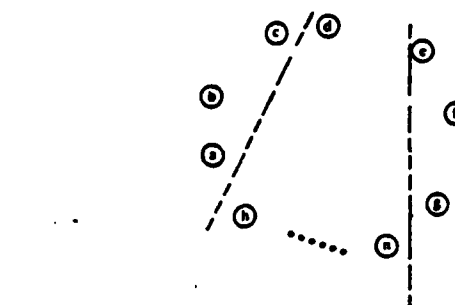


Fig. 1. Test algorithm for abstract circuit or system.

several hundred components it is reasonable to assume that at most two or three have failed simultaneously. As such, rather than solving a set of simultaneous equations in n -space the solution to our fault diagnosis problem actually lies in a two- or three-dimensional submanifold which should yield a commensurate reduction in test point requirements. Unfortunately, even though we may assume that at most two or three components have failed we do not know which two or three, and as such, some type of search is still required. Fortunately, with the aid of an appropriate decision algorithm the required search can be implemented quite simply.

Although the resultant algorithm represents a divergence from previous work by the authors and the established literature [4] we believe that it meets all of the requirements for a viable analog fault diagnosis algorithm as outlined by Pllice [8] and in [11]. In particular the algorithm:

- i) is applicable to both *linear and nonlinear systems* modeled in either the *time or frequency domain*,
- ii) can be used to locate *multiple hard or soft faults*,
- iii) and is designed to locate failures in "replaceable modules" such as an *IC chip, PC board, or subsystem* rather than in discrete components.

Moreover, this is achieved at an acceptable computational cost and with minimal test point requirements.

Consider the circuit or system which is illustrated abstractly in Fig. 1. Here, the individual chips, discrete components or subsystems are denoted by circles indexed from a to n . These components are subdivided into two groups, at each step of the test algorithm, as indicated by the dashed lines in Fig. 1. At each step we assume that one group; say, d through n ; is composed of good components and we use the known characteristics of these components together with the test data to determine whether or not the remaining components; a , b , and c in this case; are good. Of course, if components d through n are actually good

Manuscript received March 16, 1981; revised July 3, 1981 and November 15, 1981. This work was supported in part by the Joint Services Electronic Program of Texas Technological University under ONR Contract 76-C-11136.

The authors are with the Department of Electrical Engineering, Texas Technological University, Lubbock, TX 79409.

then the resultant test results for components a , b , and c will be reliable. On the other hand, if any one of the components d through n is faulty the test data on a , b , and c will be unreliable. As such, we repeat the process at the next step of the test algorithm with a different subdivision of components. For instance, we may assume that a through d and h through n are good and use their characteristics to test components e , f , and g . Finally, after a number of such repetitions the test results obtained at the various steps are analyzed to determine the faulty components.

Of course, the number of components which may be tested at any one step is dependent on the number of test points available while the number of steps required is determined by the number of components which may be tested at any one step and the bound on the maximum number of simultaneous failures. As such, the procedure yields a natural set of tradeoffs between the numbers of test points, simultaneous failures, and steps required by the algorithm. Indeed, since the computational cost associated with each step of the algorithm is essentially the cost of a single system simulation the latter parameter is a natural measure of computational cost.

In the following section we describe the simulation model used to test one set of components under the assumption that the remaining components are good. The model is formulated in both the linear and nonlinear cases and can be used as readily to test IC chips and subsystems as individual components. Moreover, the requirement that an appropriate matrix be invertible determines the maximum number of components which can be simultaneously tested from a given set of test points as well as the allowable component subdivisions. In Section III two decision algorithms for analyzing the resultant test data are described. Indeed the required theory is reminiscent, though not identical to, the t -diagnosability theory developed for digital system testing over the past decade [1], [5], [9]. In the context of our application we give an exact decision algorithm for the case of a single failure together with an analysis of the possible tradeoffs between test points and algorithm steps (read computer costs). Although an exact decision algorithm for the multifailure case has yet to be developed an heuristic algorithm which is applicable to both the single and multifailure case is presented. The algorithm, which is based on an inherently analog heuristic [6], to the effect that two analog errors will never cancel, has proven to be highly reliable while simultaneously reducing the number of steps required from that of the exact (single failure) algorithm. Finally, Section IV is devoted to a number of examples including three nonlinear circuits and five linear circuits with as many as 200 components. Using these eight circuits some 400-plus simulations of the algorithm were carried out for both the single and multiple failure cases with both hard and soft faults.

II. THE SIMULATION MODEL

Although our test algorithm can be formulated in terms of any of the standard system models for the purpose of this exposition we will assume a *component connection*

model for the circuit or system under test [3]. In the nonlinear case the *unit under test* is represented by a set of decoupled state models characterizing its components and/or subsystems together with an algebraic connection equation as follows:

$$\begin{aligned}\dot{x}_i &= f_i(x_i, a_i) \\ b_i &= g_i(x_i, a_i), \quad x_i(0) = 0, \quad i = 1, 2, \dots, n\end{aligned}\quad (2.1)$$

and

$$a = L_{11}b + L_{21}u \quad (2.2)$$

$$y = L_{21}b + L_{22}u. \quad (2.3)$$

Here, $a = \text{col}(a_i)$ is the column vector composed of the *component input variables*, $b = \text{col}(b_i)$ is the column vector composed of *component output variables*, $x_i = \text{col}(x_i)$ is the column vector composed of the *component state variables*, u is the vector of *external test inputs* applied to the system and y is the vector of *system responses measured at the various test points*. Although the component connection model is not universal it is quite general and subsumes most of the classical topological connection models commonly used in circuit and system theory [3]. Moreover, its inherently decoupled nature is ideally suited to the test problem wherein we desire to distinguish between the characteristics of the individual system components. Although these components may be taken to be elementary RLC components and/or discrete semiconductor devices, in practice the "components" are taken to be the "replaceable modules" within the circuit or system, under test; say, an IC or a "throw-away" circuit board.

At each step of the test algorithm we subdivide the "components" into two groups denoted by "1" and "2" with the components in group "1" assumed to be good and used together with the known values of u and y to compute the component input and output variables, a , and b , for the components in group "2". Although computationally we prefer to work with the decoupled component equations for notational brevity we combine the equations for the components in each group into a single equation

$$\begin{aligned}\dot{x}^1 &= f^1(x^1, a^1) \\ b^1 &= g^1(x^1, a^1), \quad x^1(0) = 0\end{aligned}\quad (2.4)$$

and

$$\begin{aligned}\dot{x}^2 &= f^2(x^2, a^2) \\ b^2 &= g^2(x^2, a^2), \quad x^2(0) = 0.\end{aligned}\quad (2.5)$$

Here, x^1 , a^1 , and b^1 are the vectors of group "1" component state variables, inputs, and outputs; and similarly for x^2 , a^2 , and b^2 . To retain notational compatibility with (2.4) and (2.5) we reorder and partition the connection equations of (2.2) and (2.3) to be conformable with (2.4) and (2.5) as follows:

$$a^1 = L_{11}^1 b^1 + L_{11}^2 b^2 + L_{12} u \quad (2.6)$$

$$a^2 = L_{21}^1 b^1 + L_{21}^2 b^2 + L_{22} u \quad (2.7)$$

$$y = L_{21}^1 b^1 + L_{21}^2 b^2 + L_{22} u. \quad (2.8)$$

Given (2.4)–(2.8) our goal is to compute the group “2” component variables, a^2 and b^2 , given the test input, u , the measured test responses, y , and an assumption to the effect that the group “1” components are not faulty. To this end we assume that L_{21}^2 admits a left inverse, $[L_{21}^2]^{-L}$, which, in turn, determines the allowable component subdivisions. Under this assumption one may then formulate a component connection model for a “pseudo circuit” composed of the group “1” components with external input vector $u^p = \text{col}(u, y)$ and external output vector $y^p = \text{col}(a^2, b^2)$ in the form

$$x^1 = f^1(x^1, a^1) \quad (2.9)$$

$$b^1 = g^1(x^1, a^1), \quad x^1(0) = 0 \quad (2.9)$$

$$a^1 = K_{11}b^1 + K_{21}u^p \quad (2.10)$$

$$y^p = K_{21}b^1 + K_{22}u^p. \quad (2.11)$$

Indeed, some algebraic manipulation of (2.6)–(2.8) together with the assumption that $[L_{21}^2]^{-L}$ exists will yield

$$K_{11} = [L_{11}^1 - L_{11}^2[L_{21}^2]^{-L}L_{21}^1] \quad (2.12)$$

$$K_{12} = [L_{12}^1 - L_{11}^2[L_{21}^2]^{-L}L_{22}^1 \mid L_{11}^2[L_{21}^2]^{-L}] \quad (2.13)$$

$$K_{21} = \begin{bmatrix} L_{21}^1 - L_{21}^2[L_{21}^2]^{-L}L_{21}^1 \\ -[L_{21}^2]^{-L}L_{21}^1 \end{bmatrix} \quad (2.14)$$

and

$$K_{22} = \begin{bmatrix} L_{12}^2 - L_{11}^2[L_{21}^2]^{-L}L_{22}^1 \mid L_{21}^2[L_{21}^2]^{-L} \\ -[L_{21}^2]^{-L}L_{22}^1 \mid [L_{21}^2]^{-L} \end{bmatrix} \quad (2.15)$$

Since, in our test problem both u and y are known, the above equations can be solved via any standard circuit analysis code to compute $y^p = (a^2, b^2)$. Now, under our assumption that the group “1” components are not faulty $y^p = (a^2, b^2)$ represents the inputs and outputs which actually appeared at the terminals of the group “2” components during the test. As such, we may determine which of the group “2” components are faulty by solving (2.5) with input a^2 and checking to determine whether or not the resultant output coincides with b^2 . Of course, since our assumption to the effect that the group “1” components are not faulty may not be valid the results of this test are not reliable. As such, we repeat the process a number of times with different choices for the subdivision of the components into group “1” and group “2”. Here, the only constraint on the choice of subdivisions is the requirement that $[L_{21}^2]^{-L}$ exist while the number of combinations employed is limited only by the cost of the required simulations. The results of the several steps in the test algorithm are then analyzed via the techniques described in the following section to determine those components which are actually faulty. To this end the results of each step of the

test algorithms are tabulated as follows:

“2” \ “1”	a	b	c	...	k
0	x				
1	y				
	⋮				
0	z				

Here a, b, c, \dots, k denote the group “1” components for a given step of the test algorithm, x, y, \dots, z denote the corresponding group “2” components while the binary annotation associated with the group “2” components indicates whether this step of the test algorithm indicated that they were good (0) or bad (1). Although this tabular notation is somewhat cumbersome we will eventually generate a binary array indexed by the group “1” and group “2” components in the process of our decision algorithm in which case the tabular notation proves to be convenient.

For linear systems one may formulate an identical algorithm in which the component equations (2.1) are modeled in the frequency domain via

$$b_i = Z_i a_i, \quad i = 1, 2, \dots, n \quad (2.16)$$

where we have suppressed the s -variable for notational brevity. Then upon subdividing the components into two groups characterized by the equalities $b^1 = Za^1$ and $b^2 = Za^2$ and solving the resultant equations under the assumption that $[L_{21}^2]^{-L}$ exists one obtains an equation in the form $y^p = Mu^p$. Specifically,

$$a^2 = M_{11}u + M_{12}y \quad (2.17)$$

$$b^2 = M_{21}u + M_{22}y \quad (2.18)$$

where

$$M_{11} = \left([L_{11}^2 - L_{11}^2[L_{21}^2]^{-L}L_{21}^1] \cdot \{1 - Z^1[L_{11}^1 - L_{12}^2[L_{21}^2]^{-L}L_{21}^1]\}^{-1} \cdot Z^1[L_{21}^1 - L_{11}^2[L_{21}^2]^{-L}L_{22}^1] + [L_{12}^2 - L_{11}^2[L_{21}^2]^{-L}L_{22}^1] \right) \quad (2.19)$$

$$M_{12} = \left([L_{11}^2 - L_{11}^2[L_{21}^2]^{-L}L_{21}^1] \cdot \{1 - Z^1[L_{11}^1 - L_{12}^2[L_{21}^2]^{-L}L_{21}^1]\}^{-1} \cdot Z^1[L_{11}^2[L_{21}^2]^{-L}] + [L_{11}^2[L_{21}^2]^{-L}] \right) \quad (2.20)$$

$$M_{21} = - \left([L_{21}^2]^{-L}L_{21}^1 \{1 - Z^1[L_{11}^1 - L_{12}^2[L_{21}^2]^{-L}L_{21}^1]\}^{-1} \cdot Z^1[L_{21}^1 - L_{11}^2[L_{21}^2]^{-L}L_{22}^1] - [L_{21}^2]^{-L}L_{22}^1 \right) \quad (2.21)$$

$$M_{22} = - \left([L_{21}^2]^{-L}L_{21}^1 \{1 - Z^1[L_{11}^1 - L_{12}^2[L_{21}^2]^{-L}L_{21}^1]\}^{-1} \cdot Z^1[L_{11}^2[L_{21}^2]^{-L}] + [L_{21}^2]^{-L} \right) \quad (2.22)$$

Although these expressions appear to be foreboding they may all be computed with the aid of only a single matrix inversion. Moreover, since the M_{ij} are independent of the test data and computed in terms of the nominal values of the group "1" components they may be computed off-line and stored in a data base to be retrieved at the time a test is conducted. Furthermore, since only a single test vector is required, single frequency testing can be employed in which case the M_{ij} need only be computed at a single frequency. As such, the only on-line computation required for the fault diagnosis of a linear system is the matrix-vector multiplication indicated by (2.17) and (2.18) together with the computation of $Z^2 a^2$.

Unlike the linear case, if one is working with a nonlinear circuit or system, the simulations required to compute a^2 and b^2 require *a priori* knowledge of y and u and thus must be carried out on-line. In practice, however, relatively few time steps are required by these simulations, thereby minimizing their running time. Moreover, all simulations are carried out with nominal components allowing one to use standard computer-aided design circuit models. Indeed, since the group "2" component models are only invoked at the final step of the analysis one can avoid simulating "troublesome" components by always including them in group "2" though this usually means that additional test points will be required. As such, one can avoid simulating "fuzzy" components which do not admit a viable simulation model and/or nonlinear components. Indeed, if sufficiently many test points are available to permit all nonlinear components to be included in group "2" a linear simulation model such as that of (2.17) and (2.18) may be employed even for a nonlinear system.

III. DECISION ALGORITHMS

Since the results of the tests described in the preceding section are dependent on our assumption that the group "1" components are not faulty they are not immediately applicable. Following the philosophy initiated by Preparata, Metze, and Chien [9] in their study of selftesting computer networks, however, if one assumes a bound on the maximum number of faulty components it is possible to determine the actual fault(s) from an analysis of the test results obtained at the various steps in the algorithm. To this end we will give a complete analysis of the theory required to locate a single fault together with an heuristic which is applicable to the multiple fault case.

Let us assume that at most one circuit component is faulty and that the test results obtained from a given step of the algorithm indicate that all group "2" components are good as indicated in the following table:

"2" \ "1"	a	b	c	...	k
0	x				
0	y				
...	...				
0	z				

In this case we claim that the group "2" components are,

in fact, good. Indeed, if a group two component were actually faulty then our test results are incorrect, which could only happen if one of the group "1" components was faulty. As such, the system would have two faulty components contradicting our assumption to the effect that at most one component is faulty.

Now, consider the case where the results from a given step of the test algorithm indicate that exactly one group "2" component is faulty; say, x

"2" \ "1"	a	b	c	...	k
1	x				
0	y				
...	...				
0	z				

In this case the same argument we used above will guarantee that the components which test good; say, y through z ; are, in fact, good. On the other hand we have no information about x . It may be faulty or, alternatively, the test result may be due to a faulty group "1" component.

Finally, consider the case where two or more group "2" components test bad in a given step as indicated in the following table:

"2" \ "1"	a	b	c	...	k
1	x				
1	y				
...	...				
0	z				

Since, under our assumption of a single failure, it is impossible for two or more group "2" components to be faulty, this test result implies that at least one of the group "1" components is bad. On the other hand, since we have assumed that there is at most one faulty component the faulty group "1" component is the only faulty component and, as such, the group "2" components are all good.

Consistent with the above, at each step of the test algorithm, either all or all but one of the group "1" components are found to be good. As such, if we choose our subdivisions so that the components which are found to be good at one step of the algorithm are included in group "1" in all succeeding steps we will eventually arrive at a group "1", all of whose components are known to be good. As such, the test results obtained at that step will be reliable, thereby allowing us to accurately determine the faulty components in group "2". Although the number of components in group "1" and group "2" may vary from step to step (especially if we work with multivariate components) if we assume that group "1" contains $n-m$ components and group "2" contains " m " components at each step of the algorithm then the process will terminate in approximately n/m steps. Since the computational cost of the algorithm is proportional to the number of steps (essentially the cost of one simulation per step) while m is determined by the number of allowable test points the ratio n/m represents a natural measure of the possible

tradeoffs between test point and computer requirements when employing the algorithm in a single fault mode.

Unlike the single fault case, at the time of this writing, we do not yet have an exact decision algorithm for the multiple fault case. Following Liu, however, the problem can be greatly simplified if one adopts an "analog heuristic" to the effect that two independent analog failures will never cancel [6]. Needless to say, this is an inherently analog heuristic since two binary failures have a fifty-fifty chance of cancelling one another. In the analog case, however, two independent failures are highly unlikely to cancel one and another (as long as one works with reasonably small tolerances). Indeed, in the totality of some 72 simulations of the algorithm using the heuristic which we have run on three different circuits it has never failed.

Recall from our discussion of the single fault case that whenever a test result indicates that a component is good then it is, in fact, good. Although this is not rigorously true in the multiple failure case it is true under the assumption of our heuristic. For instance, consider the test results indicated in the following table in which x is found to be good.

"2" \ "1"		a	b	c	\dots	k
0	x					
1	y					
\vdots	\vdots					
0	z					

Now, if x is actually faulty there must be a faulty group "1" component whose effect is to cancel the error in x as observed during this step of the test algorithm. This is, however, forbidden by our heuristic and, as such, we conclude that x is actually good.

Interestingly, our heuristic can be carried a step further than indicated above since, under our heuristic, a bad group "1" component would normally yield erroneous test results. An exception would, however, occur if some of the group "1" components are totally decoupled from some of the group "2" components. As such, if prior to our test we generate a coupling table (by simulation or a sensitivity analysis) which indicates whether or not a faulty group "1" component will effect the test results on a group "2" component, our heuristic may be used to verify that certain group "1" components are good whenever a good group "2" component is located. Consider for example the following table:

"2" \ "1"		a	b	c	\dots	k
0	x	1	0	1	-	1
1	y	1	1	0		0
\vdots	\vdots	\vdots	\vdots	\vdots		\vdots
0	z	0	1	1		0

in which a "1" in the i - j position indicates that the test results for component i are affected by component j while a "0" in the i - j position indicates that component j does

not affect the test results for component i . Now, since component x has been found to be good in this test our heuristic implies that those group "1" components which are coupled to x in this test are also good. Similarly, since z is good the heuristic implies that b and c are also good. Thus with a single test, we have verified that x , z , a , b , c , and k are all good.

Since in any practical circuit the coupling table is composed mostly of 1's it has been our experience that relatively few steps of the algorithm will yield a complete diagnosis. To implement the heuristic, however, one must assume that the maximum number of faulty components is strictly less than the number of group two components. If not, the test results at each step may show that all group "2" components are faulty, in which case no reliable test information is obtained. Moreover, the degree to which the number of group "2" components exceeds the maximum number of faulty components determines the number of algorithm steps which will be required to fully diagnose a circuit.

IV. EXAMPLES

To illustrate the exact decision algorithm for the single fault case consider a system composed of eight components: a, b, \dots, h ; in which any five may test the remaining three. Initially, we let a through e represent the group "1" components and f, g , and h represent the group "2" components and assume that the test results for this first step are as indicated in the following table:

"2" \ "1"		a	b	c	d	e
0	f					
0	g					
1	h					

Employing our exact algorithm for the single fault case the above table indicates that components f and g are good and, as such, we move them into group "1" for the second step of the algorithm obtaining

"2" \ "1"		f	g	a	b	c
0	d					
1	e					
1	h					

Since this test indicates that two group "2" components are bad which contradicts our single fault assumption the faulty component must be in group "1" implying that d , e , and h are all good. We therefore move these components into group "1" and implement the final step of our algorithm in the form

"2" \ "1"		h	e	d	f	g
0	a					
1	b					
0	c					

Since all group "1" components are known to be good this final test is reliable and indicative of the fact that b is the faulty component.

Note, that the requirement that L_{21}^2 be left invertible may make it impossible to use some component subdivisions in which case an alternative sequence of steps may be required in the above process. For instance, if h , e , d , f , and g is not an allowable subdivision the last step in the above process might be replaced by

"2"	"1"	e	d	f	g	a
1	b					
0	c					
0	h					

indicating that c and h are good. Now, a final step in which c , e , d , f , and g make up group "1" will be reliable as indicated below:

"2"	"1"	c	e	d	f	g
0	a					
1	b					
0	h					

Now, consider the same single fault example in which our heuristic algorithm is applied using the coupling table indicated below:

"2"	"1"	a	b	c	d	e
0	f	1	0	0	1	0
0	g	0	0	1	1	0
1	h	1	1	1	0	1

According to our heuristic f and g are good and, moreover, everything in group "1" which is coupled to either f or g is good. As such, we conclude from this first step that f , g , a , c , d , and e are all good. Thus taking group "1" to be e , d , f , g , and a in the next step will yield a reliable test for b , d , and h as above.

Finally, consider the case where at most two failures are assumed with the first step in our test algorithm yielding:

"2"	"1"	a	b	c	d	e
0	f	1	0	0	1	1
1	g	0	0	1	1	0
1	h	1	1	1	0	1

Consistent with our heuristic f , a , d , and e are found to be good in this step. Incorporating these components into group "1" for the following step we obtain:

"2"	"1"	f	a	c	d	e
1	b	0	1	1	1	0
1	g	1	0	1	1	0
1	h	1	1	1	0	1

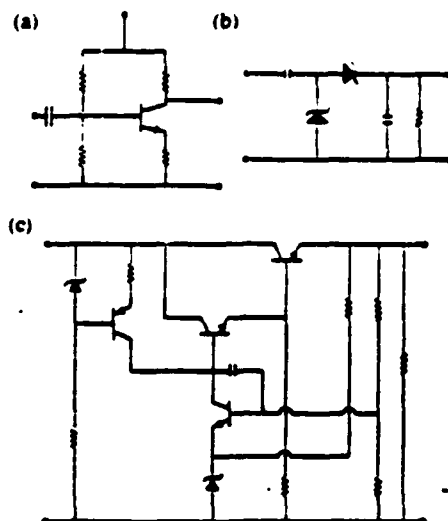


Fig. 2. Nonlinear circuits on which the proposed algorithm was tested using simulated test data.

which gives us no information in the multiple failure case.¹ As such, we try another allowable combination obtaining the following table:

"2"	"1"	f	g	a	d	e
1	b	0	1	1	1	0
0	h	1	1	1	0	1
1	c	0	1	1	1	0

indicating that h , f , g , a , and e are good. Coupled with our previous knowledge that d is good this implies that all group "1" components are good and hence this last step in our algorithm reliably indicates that b and c are the faulty components.

To obtain an estimate of the actual performance of the algorithm it was applied to 8 different circuits using simulated test data [17]. These included the 5, 6, and 14 component nonlinear circuits of Fig. 2(a)–(c), for which the full time-domain implementation of the algorithm was carried out and the linear 12, 22, 50, 100, and 200 component circuits of Fig. 3(a)–(c), on which a frequency-domain simulation was carried out. In addition multiple failure examples were run on the 12 and 22 component circuits and perturbations on the good component values were introduced into one run of the 12 component circuit to test its robustness. For the 5 and 6 component nonlinear circuits and the 12 component linear circuit the algorithm was implemented on a desk top calculator while a TI 990/20 16-bit mini and a VAX 11/780 32-bit midi were used for the larger circuits. The results of a total of 445 simulations of the algorithm are summarized in Table I. In all 445 runs the algorithm failed to give the correct answer on only four occasions. Indeed, this problem was encountered only with

¹Actually one can deduce that at least one of the group "1" components is bad since all three group "2" components cannot be bad via our two fault assumption. This, in turn, implies that at most one of the group "2" components is bad.

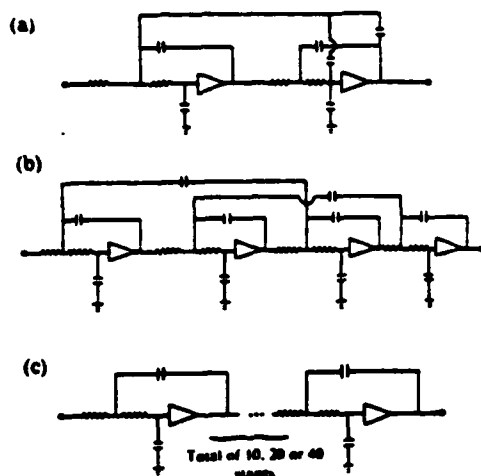


Fig. 3. Linear circuits on which the proposed algorithm was tested using simulated test data.

TABLE I

Circuit/Comp.	TP	Faults	Dec. Alg.	Avg. # Simul.	Ambiguity Set															
					1	2	3	4	5	6	7	10	20	40	80	160	320			
Nonlinear circuit 2a1 simulated in the time domain on an HP 9825	3	1	Exact	-	5	1														
	3	1	Exact	-	5															
Nonlinear circuit 2a1 simulated in the time domain on an HP 9825	2	1	Exact	-		3	2													
	2	1	Exact	-		3	2													
Nonlinear circuit 2a1 simulated in the time domain on a VAX 11/780	4	1	Exact	7.67	14															
	4	1	Exact	-	12															
Linear circuit 3a1 simulated in the time domain on an HP 9825	4	2	Exact	-		15														
	3	1	Exact	-		10	2													
	3	1	Exact	-		12														
	3	1	Exact	-		10	2													
Linear circuit 3a1 simulated in the time domain on a VAX 11/780	8	1	Exact	4.00	22															
	5	1	Exact	5.1	10				4											
	5	1	Exact	4.00	10									6						
	5	2	Exact	-	22															
	5	3	Exact	-	-	2	2													
	5	3	Exact	-	-	-	3						1			1				
	5	4	Exact	-	-	-	-	1	3	1	1									
Linear circuit 3a1 with 10 stages simulated in the time domain on a VAX 11/780	15	1	Exact	3.2	40											6			6	
	10	1	Exact	3.3	20	2											20			
Linear circuit 3a1 with 10 stages simulated in the time domain on a VAX 11/780	15	1	Exact	4.0	40												10			
Linear circuit 3a1 with 20 stages simulated in the time domain on a VAX 11/780	20	1	Exact	4.0	80													20		
Linear circuit 3a1 with 40 stages simulated in the time domain on a VAX 11/780	40	1	Exact	4.0	40														10	

the 50 component circuit of Fig. 3(c) when run on the 16-bit mini computer, wherein we believe that it was caused by the numerical error resulting from the inversion of a 50-by-50 matrix without the aid of sparse matrix techniques [17]. Indeed, the same example run on the VAX yielded an accurate diagnosis with any of the 50 components faulty. The size of the ambiguity set resulting from each simulation of the algorithm is indicated in Table I as is the mean number of on-line simulations required for the various circuits. In addition one off-line simulation per component subdivision is required to generate the coupling tables required by the heuristic decision algorithm.

V. CONCLUSIONS

Although the proposed algorithm is still new it appears to meet most of the criteria formulated by Plice [8] as well as those of [11]. Although the *on-line computational requirements* for the algorithm do not compare with a simulation-before-test algorithm they can be kept within reasonable bounds. Indeed, unlike most simulation-after-test algorithms no iterative on-line computation is required. Moreover, one can limit the on-line computation by restricting the number of algorithm steps (at the price of increasing the ambiguity in the resultant diagnosis). Furthermore, in the linear case and/or in the case where there are sufficiently many test points available to permit all nonlinear and fuzzy components to be included in group "2" the major part of the computation required by the algorithm can be done off-line.

In general the proposed algorithm permits one to trade-off between on-line computational requirements and *test points*. Indeed, as indicated in Table I, one can reduce the *test point requirements* to quite reasonable levels though this is usually achieved at the cost of increasing the number of steps in the algorithm (and hence its on-line computational requirements). In particular, our simulations indicate that the algorithm comes close to achieving the \sqrt{n} test point goal set in [11].

With respect to the remainder of the criteria specified in [11] the algorithm "looks good." In particular, all simulations are carried out using *nominal component models*, it can test *linear and nonlinear modules* of arbitrary size, and is amenable to *in situ* testing and *parallel processing* techniques (since several steps of the algorithm can be carried out simultaneously).

At the present time the major open question with respect to the performance of the algorithm is its *robustness*. Indeed, there is nothing in the algorithm to make it inherently robust, though our initial test for robustness indicated by the asterisk in Table I proved to be favorable.

REFERENCES

- [1] T. Amin, unpublished notes, Bell Labs., 1980.
- [2] H. M. S. Chen and R. Sacks, "A search algorithm for the solution of the multifrequency fault diagnosis equations," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 589-594, 1979.
- [3] R. A. DeCarlo and R. Sacks, *Interconnected Dynamical Systems*. New York: Marcel Dekker, 1981.
- [4] P. Duhamel and J. C. Rault, "Automatic test generation techniques for analog circuits and systems: A review," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 411-440, 1979.
- [5] S. L. Hakimi, "Fault analysis in digital systems—A graph theoretic approach," in *Rational Fault Analysis*, New York: Marcel Dekker, 1977, pp. 1-12.
- [6] R.-W. Liu, unpublished notes, Univ. of Notre Dame, Notre Dame, IN, 1980.
- [7] W. A. Plice, Presentation at the Workshop on Analog Fault Diagnosis, Univ. of Notre Dame, Notre Dame, IN, May 1981.
- [8] —, "A survey of analog fault diagnosis," presented at the workshop on Analog Fault Diagnosis, Univ. of Notre Dame, Notre Dame, IN, May 1981.
- [9] F. P. Preparata, G. Metze, and R. T. Chein, "On the connection assignment problem of diagnosable systems," *IEEE Trans. Electron. Comp.*, vol. EC-16, pp. 448-454, 1967.
- [10] L. M. Roytman and M. N. S. Swamy, "Multiple-fault location in analog circuits," in *Proc. IEEE Int. Symp. Circuits and Systems*, (Houston, TX), Apr. 1980, pp. 1072-1073.
- [11] R. Sacks, "Criteria for Analog Fault Diagnosis," in *Proc. European Conf. Circuit Theory and Design*, (The Hague), Aug. 1981, pp. 75-78.

- [12] R. Saeks, S. P. Singh, and R.-W. Liu, "Fault isolation via components simulation," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 634-640, 1972.
- [13] A. A. Sakla, E. I. El-Masry, and T. N. Trick, "A sensitivity algorithm for the fault detection in analog circuits," in *Proc. IEEE Int. Symp. Circuits and Systems*, (Houston), Apr. 1980, pp. 1075-1077.
- [14] A. A. Sakla and T. N. Trick, "Double fault detection in analog circuits under limited measurements," in *Proc. 12th Southeastern Symp. System Theory*, (Virginia Beach), May 1980, pp. 74-79.
- [15] T. N. Trick, W. Mayeda and A. Sakla, "Calculation of parameter values from node voltage measurements," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 466-474, 1979.
- [16] T. N. Trick, E. I. El-Masry, A. A. Sakla, and B. L. Inn, "Single fault detection in analog circuits," *Dig. 1979 Cherry Hill Test Conf.*, Cherry Hill, Oct. 1979, pp. 137-142.
- [17] C.-C. Wu, Ph.D. dissertation, Texas Tech Univ., in preparation.



Cwan-chia Wu was born in Tainan, Taiwan, Republic of China, in 1955. He received the B.S. and M.S. degrees in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 1977 and 1979, respectively. He received the Ph.D. degree in 1981.

From 1979 to 1981, he was a Research Assistant of Electrical Engineering, Texas Tech University, Lubbock, TX. From 1982 he has been National Taiwan Institute of Technology, Taipei, Taiwan. His present interests are in the area of

fault analysis and microprocessor applications.



Kazuo Nakajima (S'79-M'80) was born in Nishinomiya, Hyogo, Japan, on April 2, 1950. He received the B. Eng. and M. Eng. degrees, both in control engineering, from Osaka University, Toyonaka, Osaka, Japan, in 1973 and 1975, respectively, and the Ph.D. degree in computer science from Northwestern University, Evanston, IL, in 1980.

He was a Research Assistant from 1978 to 1979 and a Post-Doctoral Fellow from 1979 to 1980, both in computer science at Northwestern

University. He has been an Assistant Professor of Computer Science and Electrical Engineering at Texas Tech University since August 1980. His research interests include system diagnosis, scheduling theory, analysis of algorithms, computational complexity, graph theory, and VLSI algorithms.

Dr. Nakajima is a member of ACM and Institute of Electronics and Communication Engineers of Japan.



Chin-Long Wey received the B.S. degree in mathematics from the National Central University, Chungli, Taiwan, and the M.S. in mathematics and computer science at Texas Tech University, Lubbock, TX. He is presently working towards the Doctorate degree in electrical engineering at Texas Tech with a dissertation on the analog fault diagnosis problem.



Richard Saeks (S'59-M'65-SM'74-F'77) was born in Chicago, IL, in 1941. He received the B.S. degree in 1964 from Northwestern University, Evanston, IL, the M.S. degree in 1965 from Colorado State University, Fort Collins, and the Ph.D. degree in 1967 from Cornell University, all in electrical engineering.

He is presently Paul Whitfield Horn Professor of electrical engineering, mathematics, and computer science at Texas Technological University, Lubbock, where he is involved in teaching and research in the areas of fault analysis, large-scale systems, and mathematical system theory.

Dr. Saeks is a member of AMS, SIAM, ACM, ASEE, and Sigma Xi.

MULTIDIMENSIONAL SYSTEM THEORY

205

286

FEEDBACK SYSTEM DESIGN:

The Single-Variate Case — Part I*

R. Saeks¹, J. Murray¹, O. Chua², C. Karmokolias³,
and A. Iyer¹

Abstract. A recently developed algebraic approach to the feedback system design problem is reviewed via the derivation of the theory in the single-variate case. This allows the simple algebraic nature of the theory to be brought to the fore while simultaneously minimizing the complexities of the presentation. Rather than simply giving a single solution to the prescribed design problem we endeavor to give a complete parameterization of the set of compensators which meet specifications. Although this might at first seem to complicate our theory it, in fact, opens the way for a sequential approach to the design problem in which one parameterizes the subset of those compensators which meet the second specification...etc. Specific problems investigated include feedback system stabilization, the tracking and disturbance rejection problem, robust design, transfer function design, pole placement, simultaneous stabilization, and stable stabilization.

1. Introduction

In 1976 Youla, Bongiorno, and Jabr published two, now classical, papers [23,24] in which a complete parameterization of the set of stabilizing compensators for a multivariate feedback system was obtained. In the ensuing years this work, which is often termed the YBJ theory, has led to the development of an entirely new approach to the *feedback system design problem*. Indeed, their stabilization theory has been extended to include:

- (i) the tracking and disturbance rejection problem
- (ii) robust design algorithms
- (iii) design with a proper or stable compensator
- (iv) transfer function design

*Received September 16, 1981; revised November 10, 1981. This research was supported by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

¹Department of Electrical Engineering, Texas Tech University, Lubbock, Texas 79409, USA.

²Presently with Honeywell, Inc., Phoenix, Arizona.

³Presently with Kearfott Division, Singer, Inc., Little Falls, New Jersey.

- (v) pole placement
- (vi) simultaneous stabilization

Moreover, much of the work has been extended to the case of general linear systems; distributed, time-varying, multidimensional, etc.; by formulating it in an abstract ring theoretic [8,10] or algebro-geometric setting [17]. Unfortunately, these generalizations have been achieved at the cost of increasing the complexity of the theory and, as such, the simple algebraic character of the work has been obscured.

The purpose of the present paper is to survey this literature in such a way as to illustrate the simplicity of the theory. To this end the presentation is restricted to the single-variate case wherein a simple algebraic theory is possible. Indeed, by so doing we are able to give simple single-variate algebraic derivations for several results whose true character has hitherto been obscured by the abstract ring theoretic or multivariable theory.

The key to our theory is a three step design philosophy

- (i) stabilization
- (ii) achievement of design constraints
- (iii) optimization of system performance

First and foremost, a feedback system must be *stable* and, as such, the first step in the design process is the *parameterization of all stabilizing compensators* for the given plant. Although it might suffice to specify a single stabilizing compensator if our goal was simply to design a stable system, in practice stabilization is only the first step of the design process. As such, we must characterize all stabilizing compensators if we are to choose among the stabilizing compensators to find one which also achieves the design constraints and/or optimizes system performance. A complete parameterization of the set of stabilizing compensators for the given plant is thus obtained as a first step in the design process. Indeed, the parameterization is chosen in such a way that the various feedback system gains are linear (affine) in the resultant design parameter, thereby setting the stage for the choice of a design parameter which also achieves the *design constraints* and/or *optimizes some measure of system performance*.

Once the stabilizing compensators have been characterized, step two of the design process is to choose a subset of the stabilizing compensators which also achieve the prescribed design constraints; tracking and disturbance rejection, transfer function specification, robustness, etc. Finally, if any remaining design latitude exists after the design constraints have been met it may be used to optimize some measure of system performance; sensitivity, energy consumption, etc.

The paper is divided into two parts dealing with the classical *asymptotic design problems*: stabilization, tracking, and disturbance rejection; and a survey of modern frequency domain design; robust design, pole placement, simultaneous design, respectively. In the remainder of this introduction the

fractional representation theory for a single-variate system is developed. The key to this theory lies with the representation of a rational function as a ratio of stable rational functions rather than as a ratio of polynomials. Such a formulation opens the door to the desired generalization, wherein stability is well defined even though no analog of a polynomial exists. Moreover, it yields what we believe to be a more *natural concept of coprimeness* in which only cancellations between (closed) right half-plane zeros are forbidden. Indeed, the (strict) left half-plane plays only a minimal role in the theory.

In Section 2 a derivation of the *YBJ stabilization theory* is formulated in terms of a stable coprime fractional representation. Although this derivation has appeared before [4,8], even in the single-variate case, a complete proof is given because of its fundamental nature to the remainder of the work. Indeed, the proof technique introduced here is repeated, in one form or another, throughout the paper.

Sections 3-5 are devoted to the *tracking and disturbance rejection problems* [4]. Unlike the stabilization problem a solution to these problems may fail to exist. Necessary and sufficient conditions for the existence of a solution are, however, given in the form of appropriate coprimeness criteria and a complete parameterization of the required set of compensators is obtained when these criteria are satisfied.

Part II of the paper begins with Section 6 in which the problem of *robust design* is taken up. Unlike the stabilization problem for which every solution is robust a solution to the tracking and/or disturbance rejection problem may fail to be robust. Surprisingly, however, whenever these problems are solvable they are robustly solvable. As such, beginning with the same coprimeness criteria used in the non-robust case we give an explicit parameterization for the set of compensators which robustly solve the tracking and disturbance rejection problems. This result, however, only applies to our single-variate case. In the general multivariate case a robust solution may fail to exist even though a non-robust solution exists [11].

In Section 7 the problem of designing a compensator which simultaneously stabilizes a feedback system and realizes a prescribed input-output feedback system gain is investigated. The required existence criteria for this *transfer function design problem* are formulated in terms of a divisibility condition in the ring of stable transfer functions. This is followed in Section 8 by an investigation of the *pole placement problem* in which one desires to construct a compensator which will simultaneously stabilize a system and place the poles of its input-output gain at prescribed points in the left half-plane. Interestingly, the extent to which this end can be achieved is determined precisely by the "degree" to which the plant fails to be miniphase.

In Section 9 the problem of designing a compensator which simultaneously stabilizes two distinct plants is solved. Although this *two plant problem* is a very special case of the general *simultaneous design problem* it

is the one example of the problem for which a definitive frequency domain design criterion exists [7] and is thus indicative of the direction of future research in this area. Moreover, the problem of stabilizing a feedback system with a *stable compensator* [25] proves to be a special case of this two plant problem which is developed in Section 10.

Section 11 is devoted to a short discussion of the "*optimization problem*" associated with step three of our design process. Since the specific optimization one might choose to undertake is dependent on the physical system under study and its application the development in this section concentrates on the interface between our theory and the optimization process, without going into specifics.

Finally, Section 12, is devoted to an *historical overview* of the theory and a discussion of the various *generalizations and extensions* which thus far have been formulated.

Our system will be described by a rational function

$$r(s) = \frac{p(s)}{q(s)} \quad (1.1)$$

Such a system is said to be *stable* if its poles lie in the (strict) left half-plane. Since the point at infinity is taken to lie on the imaginary axis this implies that $r(s)$ is stable if and only if it is a proper rational function and $q(s)$ is a (strictly) Hurwitz polynomial.

A *fractional representation* for $r(s)$ is a factorization of $r(s)$ in the form

$$r(s) = \frac{n_r(s)}{d_r(s)} \quad (1.2)$$

where both $n_r(s)$ and $d_r(s)$ are stable and $d_r(s) \neq 0$. If one is given a *polynomial fractional representation* for $r(s)$ such as in equation 1.1 then one can take

$$n_r(s) = \frac{p(s)}{m(s)} \quad (1.3)$$

and

$$d_r(s) = \frac{q(s)}{m(s)} \quad (1.4)$$

where $m(s)$ is any Hurwitz polynomial such that the order of $m(s)$ equals the order of $r(s)$ verifying the existence of the required fractional representation [11].

We say that the fractional representation $r(s) = n_r(s)/d_r(s)$ is *coprime* if there exist stable rational functions, $u_r(s)$ and $v_r(s)$, such that

$$u_r(s)n_r(s) + v_r(s)d_r(s) = 1 \quad (1.5)$$

Recall that for polynomials 1.5 is equivalent to the requirement that $n_r(s)$ and $d_r(s)$ do not have any common zeros [2]. In our case, however, where we are dealing with stable rational functions, Equation 1.5 implies that $n_r(s)$ and $d_r(s)$ have no common (closed) right half-plane zeros and conversely [4,11]. Although this represents a departure from classical control theory only right half-plane zeros cause instability and, as such, it is appropriate that only right half-plane pole zero cancellations be forbidden.

Unlike the classical polynomial fractional representation theory wherein the units are the constant functions in our theory the units are the *miniphase* rational functions which are stable and admit a stable inverse. That is, if $r(s) = p(s)/q(s)$ then $p(s)$ and $q(s)$ are both Hurwitz polynomials of the same order. As such, the classical theorems for polynomial fractional representations may be reformulated in our setting with the units taken to be miniphase rational functions as follows.

1. Property. Let $r(s) = n_r(s)/d_r(s)$ be coprime fractional representation for $r(s)$ and assume that $n_r(s)$ and $d_r(s)$ admit a common divisor, $k(s)$, such that

$$d_r(s) = y_r(s)k(s)$$

and

$$n_r(s) = x_r(s)k(s)$$

where $y_r(s)$, $x_r(s)$ and $k(s)$ are stable. Then $k(s)$ is miniphase. That is, the only *common divisor* of a coprime fractional representation is a unit.

Proof. Since $d_r(s) = y_r(s)k(s)$ and $n_r(s) = x_r(s)k(s)$ are coprime there exist stable $u_r(s)$ and $v_r(s)$ such that

$$1 = u_r(s)n_r(s) + v_r(s)d_r(s) = [u_r(s)x_r(s) + v_r(s)y_r(s)]k(s) \quad (1.6)$$

showing that $[u_r(s)x_r(s) + v_r(s)y_r(s)]$ is a stable inverse for $k(s)$ and hence verifying that $k(s)$ is miniphase.

2. Property. Let $r(s) = n(s)/d(s)$ be a fractional representation for $r(s)$ and let $r(s) = x_r(s)/y_r(s)$ be a coprime fractional representation for $r(s)$. Then there exists a stable $k(s)$ such that

$$n_r(s) = x_r(s)k(s)$$

and

$$d_r(s) = y_r(s)k(s)$$

Proof. Given the two fractional representations let $k(s) = d_r(s)/y_r(s)$. Then clearly $d_r(s) = y_r(s)k(s)$ while

$$n(s) = \frac{n_r(s)}{d_r(s)} d_r(s) = r(s)d(s) = \frac{x_r(s)}{y_r(s)} d_r(s) = x_r(s)k(s) \quad (1.7)$$

showing that $k(s)$ is a common factor of $n_r(s)$ and $d_r(s)$. It thus remains to verify that $k(s)$ is stable. To this end recall that since $x_r(s)/y_r(s)$ is coprime there exists stable $u_r(s)$ and $v_r(s)$ such that

$$[u_r(s)x_r(s) + v_r(s)y_r(s)] = 1 \quad (1.8)$$

hence

$$\begin{aligned} k(s) &= \frac{d_r(s)}{y_r(s)} = [u_r(s)x_r(s) + v_r(s)y_r(s)] \frac{d_r(s)}{y_r(s)} \\ &= \frac{u_r(s)x_r(s)d_r(s)}{y_r(s)} + v_r(s)d_r(s) = u_r(s)r(s)d_r(s) + v_r(s)d_r(s) \quad (1.9) \\ &= \frac{u_r(s)n_r(s)d_r(s)}{d_r(s)} + v_r(s)d_r(s) = u_r(s)n_r(s) + v_r(s)d_r(s) \end{aligned}$$

showing that $k(s)$ is stable since we have expressed it as a sum of products of stable rational functions.

Note that since a coprime fractional representation always exists [11] for $r(s)$ Property 2 implies that any pair of stable rational functions, $x_r(s)$ and $y_r(s)$, can be expressed in the form $x_r(s) = n_r(s)k(s)$ and $y_r(s) = d_r(s)k(s)$ where $n_r(s)$ and $d_r(s)$ are coprime stable rational functions and $k(s)$ is stable. As such, $k(s)$ represents a *greatest common divisor* for $x_r(s)$ and $y_r(s)$ which is unique up to a miniphase factor via Property 1.

3. Property. Let $r(s) = n_r(s)/d_r(s)$ be a coprime fractional representation for $r(s)$. Then $r(s)$ is stable if and only if $d_r(s)$ is miniphase.

Proof. If $d_r(s)$ is miniphase then $1/d_r(s)$ is stable and hence $r(s)$, being the product of two stable functions is stable. Conversely, if $r(s)$ is stable then we may express $n_r(s)$ and $d_r(s)$ in the form

$$n_r(s) = r(s)d_r(s) \quad (1.10)$$

and

$$d_r(s) = 1 d_r(s) \quad (1.11)$$

showing that $d_r(s)$ is a common factor of the coprime rational functions $n_r(s)$ and $d_r(s)$. As such Property 1 implies that $d_r(s)$ is miniphase as was to be shown.

4. Example. Consider the rational function

$$r(s) = \frac{(s+1)}{(s^2-4)} = \frac{\frac{(s+1)}{(s+1)^2}}{\frac{(s-2)}{(s+2)}} = \frac{n_r(s)}{d_r(s)} \quad (1.12)$$

FEEDBACK SYSTEM DESIGN

Now, $n_r(s)$ has zeros at $s = -1$ and $s = \infty$ while $d_r(s)$ has a zero at $s = 2$. As such, $n_r(s)$ and $d_r(s)$ have no common right half-plane zeros thus implying the existence of stable rational functions $u_r(s)$ and $v_r(s)$ such that $[u_r(s)n_r(s) + v_r(s)d_r(s)] = 1$. Indeed,

$$\left[\frac{16}{3} \right] \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right] \left[\frac{(s-2)}{(s+2)} \right] = u_r(s)n_r(s) + v_r(s)d_r(s) = 1 \quad (1.13)$$

Note that unlike the case of a polynomial fractional representation the existence of common (strict) left half-plane zeros does not preclude coprimeness. Indeed, an alternative coprime fractional representation for the above rational function takes the form

$$r(s) = \left[\frac{(s+1)}{(s^2-4)} \right] = \frac{\left[\frac{(s+1)^2}{(s+2)^3} \right]}{\left[\frac{(s-2)(s+1)}{(s+2)^2} \right]} = \frac{n'_r(s)}{d'_r(s)} \quad (1.14)$$

where $n'_r(s)$ and $d'_r(s)$ have a common zero at $s = -1$. There are, however, still coprime since

$$\left[\frac{16(s+2)}{3(s+1)} \right] \left[\frac{(s+1)^2}{(s+2)^3} \right] + \left[\frac{(s+2/3)}{(s+1)} \right] \left[\frac{(s-2)(s+1)}{(s+2)^2} \right] = u'_r(s)n'_r(s) + v'_r(s)d'_r(s) = 1 \quad (1.15)$$

2. Stabilization

The basic feedback system with which we deal is shown in Figure 1.

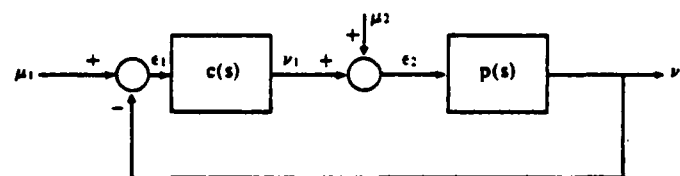


Figure 1. Basic feedback system.

For this system the usual algebraic manipulations [8] will yield the *feedback system gains*

$$\begin{bmatrix} \epsilon_1(s) \\ \epsilon_2(s) \end{bmatrix} = \begin{bmatrix} h_{\epsilon_1\mu_1}(s) & h_{\epsilon_1\mu_2}(s) \\ h_{\epsilon_2\mu_1}(s) & h_{\epsilon_2\mu_2}(s) \end{bmatrix} \begin{bmatrix} \mu_1(s) \\ \mu_2(s) \end{bmatrix} \quad (2.1)$$

where

$$\begin{bmatrix} h_{e_1\mu_1}(s) & h_{e_1\mu_2}(s) \\ h_{e_2\mu_1}(s) & h_{e_2\mu_2}(s) \end{bmatrix} = \begin{bmatrix} \frac{1}{1+p(s)c(s)} & \frac{-p(s)}{1+p(s)c(s)} \\ \frac{c(s)}{1+p(s)c(s)} & \frac{1}{1+p(s)c(s)} \end{bmatrix} \quad (2.2)$$

while

$$\begin{bmatrix} v_1(s) \\ v_2(s) \end{bmatrix} = \begin{bmatrix} h_{v_1\mu_1}(s) & h_{v_1\mu_2}(s) \\ h_{v_2\mu_1}(s) & h_{v_2\mu_2}(s) \end{bmatrix} \begin{bmatrix} \mu_1(s) \\ \mu_2(s) \end{bmatrix} \quad (2.3)$$

where

$$\begin{aligned} \begin{bmatrix} h_{v_1\mu_1}(s) & h_{v_1\mu_2}(s) \\ h_{v_2\mu_1}(s) & h_{v_2\mu_2}(s) \end{bmatrix} &= \begin{bmatrix} h_{e_2\mu_1}(s) & h_{e_2\mu_2}(s) - 1 \\ 1 - h_{e_1\mu_1}(s) & -h_{e_1\mu_2}(s) \end{bmatrix} \\ &= \begin{bmatrix} \frac{c(s)}{1+p(s)c(s)} & \frac{-p(s)c(s)}{1+p(s)c(s)} \\ \frac{p(s)c(s)}{1+p(s)c(s)} & \frac{p(s)}{1+p(s)c(s)} \end{bmatrix} \end{aligned} \quad (2.4)$$

Of course, the system is said to be stable if each of the eight feedback system gains of equations 2.1 through 2.4 is stable. Since the input/output gains, h_{μ_i} , are expressed in terms of the input/plant-input gains, h_{e_i} , via equation 2.4 this will be this case if and only if the input/plant-input gains are all stable.

For our stabilization theory we assume that a coprime fractional representation for the plant is given in the form

$$p(s) = \frac{n_p(s)}{d_p(s)} \quad (2.5)$$

where $n_p(s)$ and $d_p(s)$ are stable, $d_p(s)$ is not identically zero, and there exists stable $u_p(s)$ and $v_p(s)$ such that

$$u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (2.6)$$

In our single-variate setting every plants admits such a representation and hence we may assume 2.5 and 2.6 without loss of generality. Our goal is to characterize the set of compensators, represented in the form

$$c(s) = \frac{n_c(s)}{d_c(s)} \quad (2.7)$$

where $n_c(s)$ and $d_c(s)$ are stable and $d_c(s)$ is not identically zero, which stabilize the feedback system. Of course, we would also like 2.7 to be a coprime fractional representation. Indeed, so as to prevent (right half-plane)

pole-zero cancellation between $p(s)$ and $c(s)$ we require the stronger condition that

$$p(s)c(s) = \frac{n_p(s)n_c(s)}{d_p(s)d_c(s)} \quad (2.8)$$

be a coprime fractional representation.

Substituting the fractional representations $p(s) = n_p(s)/d_p(s)$ and $c(s) = n_c(s)/d_c(s)$ into 2.2 and 2.4 now yields

$$\begin{bmatrix} h_{1\mu_1}(s) & h_{1\mu_2}(s) \\ h_{2\mu_1}(s) & h_{2\mu_2}(s) \end{bmatrix} = \begin{bmatrix} \frac{d_p(s)d_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} & \frac{-n_p(s)d_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} \\ \frac{d_p(s)n_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} & \frac{d_p(s)d_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} \end{bmatrix} \quad (2.9)$$

and

$$\begin{bmatrix} h_{1\nu_1}(s) & h_{1\nu_2}(s) \\ h_{2\nu_1}(s) & h_{2\nu_2}(s) \end{bmatrix} = \begin{bmatrix} \frac{d_p(s)n_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} & \frac{-n_p(s)n_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} \\ \frac{n_p(s)n_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} & \frac{n_p(s)d_c(s)}{d_p(s)d_c(s) + n_p(s)n_c(s)} \end{bmatrix} \quad (2.10)$$

Since the fractional representation for $p(s)c(s)$ given in 2.8 is coprime there exist stable $p(s)$ and $q(s)$ such that

$$p(s)[d_p(s)d_c(s)] + q(s)[n_p(s)n_c(s)] = 1 \quad (2.11)$$

hence

$$[p(s) - q(s)][d_p(s)d_c(s)] + q(s)[d_p(s)d_c(s) + n_p(s)n_c(s)] = 1 \quad (2.12)$$

showing that the fractional representation for $h_{1\mu_1}(s)$ given in 2.9 is coprime. As such, it follows from Property 1 that $h_{1\mu_1}(s)$ is stable if and only if the common denominator, $[d_p(s)d_c(s) + n_p(s)n_c(s)]$ is miniphase. Moreover, since $h_{1\mu_1}(s)$ must be stable for the feedback system to be stable it follows that $[d_p(s)d_c(s) + n_p(s)n_c(s)]$ must be miniphase for the system to be stable. Conversely, if this common denominator is miniphase the system is clearly stable via 2.9 and 2.10. Moreover, if the common denominator is miniphase

$$\begin{aligned} & [d_p(s)d_c(s) + n_p(s)n_c(s)]^{-1} [d_p(s)d_c(s)] \\ & + [d_p(s)d_c(s) + n_p(s)n_c(s)]^{-1} [n_p(s)n_c(s)] = 1 \end{aligned} \quad (2.13)$$

showing that the corresponding fractional representation for $p(s)c(s)$ is coprime. We have therefore proven the following.

5. Property. Let $p(s) = n_p(s)/d_p(s)$ be a coprime fractional representation for $p(s)$ and let $c(s) = n_c(s)/d_c(s)$ be a fractional representation for $c(s)$. Then the feedback system is stable and $p(s)c(s) = [n_p(s)n_c(s)]/[d_p(s)d_c(s)]$ is coprime if and only if $[d_p(s)d_c(s) + n_p(s)n_c(s)]$ is miniphase.

Consistent with Property 5 the goal of the feedback system stabilization problem is to characterize the compensators, $c(s) = n_c(s)/d_c(s)$, such that $[d_p(s)d_c(s) + n_p(s)n_c(s)]$ is miniphase given the coprime fractional representation

$$p(s) = \frac{n_p(s)}{d_p(s)} \quad (2.14)$$

where

$$u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (2.15)$$

for some stable $u_p(s)$ and $v_p(s)$.

Stabilization Theorem: For the feedback system of Figure 1 let the plant have a coprime fractional representation as per equation 2.14 and 2.15. Then for any stable $w(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero the compensator

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]} = \frac{n_c(s)}{d_c(s)}$$

stabilizes the feedback system and yields a coprime fractional representation on $p(s)c(s) = [n_p(s)n_c(s)]/[d_p(s)d_c(s)]$. Conversely, every such stabilizing compensator is of this form for some stable $w(s)$.

Proof. According to Property 5 it suffices to characterize the class of stable $n_c(s)$ and $d_c(s)$ such that

$$d_p(s)d_c(s) + n_p(s)n_c(s) = k(s) \quad (2.16)$$

where $k(s)$ is an arbitrary miniphase function. To this end we will attempt to compute all possible stable solutions to equation 2.16. Multiplying equation 2.15 through by $k(s)$ yields

$$[k(s)u_p(s)]n_p(s) + [k(s)v_p(s)]d_p(s) = k(s) \quad (2.17)$$

verifying that

$$n_c(s) = k(s)u_p(s) \quad (2.18)$$

and

$$d_c(s) = k(s)v_p(s) \quad (2.19)$$

are particular solutions to equation 2.16. On the other hand

FEEDBACK SYSTEM DESIGN

$$d_p(s) [n_p(s)r(s)] + n_p(s) [-d_p(s)r(s)] = 0 \quad (2.20)$$

for all stable $r(s)$ showing that

$$n_c^h(s) = -d_p(s)r(s) \quad (2.21)$$

and

$$d_c^h(s) = n_p(s)r(s) \quad (2.22)$$

are homogeneous solutions to 2.16 for all stable $r(s)$. It remains to show that 2.21 and 2.22 represent all homogeneous solutions. To this end let $\underline{n}_c^h(s)$ and $\underline{d}_c^h(s)$ represent arbitrary stable homogeneous solutions to 2.16. That is

$$d_p(s)\underline{d}_c^h(s) + n_p(s)\underline{n}_c^h(s) = 0 \quad (2.23)$$

in which case we will show that they take the form of 2.21 and 2.22. As a candidate for $r(s)$ let us take $r(s) = -\underline{n}_c^h(s) / d_p(s)$ in which case we have

$$\underline{n}_c^h(s) = -d_p(s)r(s) \quad (2.24)$$

verifying 2.21 and

$$\underline{d}_c^h(s) = \frac{-n_p(s)\underline{n}_c^h(s)}{d_p(s)} = n_p(s)r(s) \quad (2.25)$$

verifying 2.22. It thus remains to show that $r(s) = -\underline{n}_c^h(s) / d_p(s)$ is stable for which we have

$$\begin{aligned} r(s) &= -\frac{\underline{n}_c^h(s)}{d_p(s)} = -\frac{\underline{n}_c^h(s)}{d_p(s)} [u_p(s)n_p(s) + v_p(s)d_p(s)] \\ &= -\frac{\underline{n}_c^h(s)n_p(s)u_p(s)}{d_p(s)} - \underline{n}_c^h(s)v_p(s) = \frac{d_p(s)\underline{d}_c^h(s)u_p(s)}{d_p(s)} - \underline{n}_c^h(s)v_p(s) \\ &= \underline{d}_c^h(s)u_p(s) - \underline{n}_c^h(s)v_p(s) \end{aligned} \quad (2.26)$$

showing that $r(s)$ is stable since it is expressed as the sum of products of stable rational functions. As such, the entire solution space for equation 2.16 takes the form

$$n_c(s) = n_c^h(s) + n_c^p(s) = -r(s)d_p(s) + k(s)u_p(s) \quad (2.27)$$

and

$$d_c(s) = d_c^h(s) + d_c^p(s) = r(s)n_p(s) + k(s)v_p(s) \quad (2.28)$$

where $k(s)$ is an arbitrary miniphase function and $r(s)$ is an arbitrary stable function.

Now, assuming that $r(s)$ and $k(s)$ are chosen so that $c_c(s)$ is not identically zero we obtain the desired set of compensators in the form

$$c(s) = \frac{[-r(s)d_p(s) + k(s)u_p(s)]}{[r(s)n_p(s) + k(s)v_p(s)]}$$

$$= \frac{[-r(s)/k(s)]d_p(s) + u_p(s)}{[r(s)/k(s)]n_p(s) + v_p(s)} = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]} \quad (2.29)$$

where $w(s) = r(s)/k(s)$ spans the set of stable rational functions such that $[w(s)n_p(s) + v_p(s)]$ is not identically zero.

In addition to giving a complete parameterization of the stabilizing compensators if one views $w(s)$ rather than $c(s)$ as the underlying design parameter for our feedback system the expressions for the various feedback system gains are greatly simplified. This follows by observing that the compensator of the theorem yields the common denominator

$$[d_p(s)d_c(s) + n_p(s)n_c(s)] = 1 \quad (2.30)$$

(since we have divided $k(s)$ out of the expression for $c(s)$). As such, the denominators in Equations 2.9 and 2.10 drop out yielding the following expressions for the feedback system gains which are linear (actually affine) in the design parameter $w(s)$.

6. Corollary. The feedback system gains which result from the use of the compensator of the stabilization theorem take the form

$$\begin{bmatrix} h_{11}(s) & h_{12}(s) \\ h_{21}(s) & h_{22}(s) \end{bmatrix} = \begin{bmatrix} w(s)n_p(s)d_p(s) + v_p(s)d_p(s) & -w(s)n_p^2(s) - v_p(s)n_p(s) \\ -w(s)d_p^2(s) + u_p(s)d_p(s) & w(s)n_p(s)d_p(s) + v_p(s)d_p(s) \end{bmatrix}$$

and

$$\begin{bmatrix} h_{11}(s) & h_{12}(s) \\ h_{21}(s) & h_{22}(s) \end{bmatrix} = \begin{bmatrix} -w(s)d_p^2(s) + u_p(s)d_p(s) & w(s)n_p(s)d_p(s) - u_p(s)n_p(s) \\ -w(s)n_p(s)d_p(s) + u_p(s)n_p(s) & w(s)n_p^2(s) + v_p(s)n_p(s) \end{bmatrix}$$

Proof. These relationships result immediately upon substituting 2.30 and the expressions for $n_c(s)$ and $d_c(s)$ of the theorem into equations 2.9 and 2.10.

The theorem gives a complete parameterization of the stabilizing compensators for our feedback system modulo the requirement that $w(s)n_p(s) + v_p(s)$ not be identically zero. Needless to say, in our single-variate case this requirement is trivially verified. Moreover, $w(s)n_p(s) + v_p(s)$ is always non-zero for some $w(s)$ hence the existence of

a stabilizing compensator for any single-variate plant is guaranteed.

7. Corollary. Every single-variate plant admits a stabilizing compensator.

Proof: Consistent with the theorem it suffices to verify the existence of a stable $w(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. Indeed, either $w(s) = -1$ or $w(s) = 0$ suffices. If $w(s) = -1$ fails, i.e.,

$$d_c(s) = -n_p(s) + v_p(s) = 0 \quad (2.31)$$

then the coprimeness equality

$$1 = u_p(s)n_p(s) + v_p(s)d_p(s) = [u_p(s) + d_p(s)]v_p(s) \quad (2.32)$$

implies that $v_p(s)$ is miniphase, since $[u_p(s) + d_p(s)]$ is a stable inverse for $v_p(s)$, in which case

$$d_c(s) = [0]n_p(s) + v_p(s) = v_p(s) \quad (2.33)$$

is not identically zero.

Note that the above result is contingent on the existence of a coprime fractional representation for the plant and therefore may fail in the various generalized settings to which the theory can be extended [8]. It does, however, hold in the multivariate case wherein a coprime fractional representation is also assured to exist [11].

Occasionally, one desires to design a compensator which is a proper rational function; i.e., $c(\infty) < \infty$; rather than simply asking for a stabilizing compensator [2, 11]. Now, $c(s) = n_c(s)/d_c(s)$ is coprime via Equation 2.30 hence $n_c(\infty)$ and $d_c(\infty)$ are not both simultaneously zero. On the other hand $n_c(\infty) < \infty$ since $n_c(s)$ is stable hence

$$c(\infty) = \frac{n_c(\infty)}{d_c(\infty)} < \infty \quad (2.34)$$

if and only if $d_c(\infty) = w(\infty)n_p(\infty) + v_p(\infty) \neq 0$. Of course, in this case $w(s)n_p(s) + v_p(s)$ is not identically zero showing that the proper stabilizing compensators take the form

$$c(s) = \frac{-w(s)d_p(s) + u_p(s)}{w(s)n_p(s) + v_p(s)} \quad (2.35)$$

where $w(s)$ is stable and

$$w(\infty)n_p(\infty) + v_p(\infty) \neq 0 \quad (2.36)$$

We may now consider two cases. First, if the plant is strictly proper, $p(\infty) = 0$, then $n_p(\infty) = 0$ and since $n_p(s)$ and $v_p(s)$ are coprime via 2.15 implies that $v_p(\infty) \neq 0$. As such, 2.36 is satisfied for all stable $w(s)$. On the other hand, if the plant is not strictly proper, $p(\infty) \neq 0$, then $n_p(\infty) \neq 0$ in which case 2.36 reduces to $w(\infty) \neq -v_p(\infty)/n_p(\infty)$. We have thus verified the following corollaries [10, 11].

8. Corollary. If $p(s)$ is strictly proper then the set of compensators given by the theorem are all proper and well defined for every stable $w(s)$.

9. Corollary. If $p(s)$ is not strictly proper then the compensators given by the theorem are well defined and proper if and only if

$$w(\infty) = \frac{-v_p(\infty)}{n_p(\infty)}$$

Finally, rather than simply looking for a proper compensator we may desire to design a stable compensator [25]. Although such a compensator does not, in general, exist, a criterion for the existence of a stable stabilizing compensator and an algorithm for its construction is given in Section 10 as a corollary to the simultaneous stabilization theorem. The result is, however, far from elementary and no parameterization of the space of such compensators is known [17].

10. Example. For the plant of Example 4 with the coprime fractional representation of of Equations 1.12 and 1.13 the required set of stabilizing compensators take the form

$$c(s) = \frac{-w(s) \left[\frac{(s-2)}{(s+2)} \right] + \left[\frac{16}{3} \right]}{w(s) \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right]} \quad (2.37)$$

with

$$c(\infty) = -w(\infty) + \frac{16}{3} < \infty \quad (2.38)$$

verifying that the resultant compensator is, indeed, proper given a strictly proper plant.

Now, let us repeat the above example using the alternative coprime fractional representation of Equation 1.14 and 1.15 which yields

$$\begin{aligned} c(s) &= \frac{-w(s) \left[\frac{(s-2)(s+1)}{(s+2)^2} \right] + \left[\frac{16(s+2)}{3(s+1)} \right]}{w(s) \left[\frac{(s+1)^2}{(s+2)^3} \right] + \left[\frac{(s+2/3)}{(s+1)} \right]} \\ &= \frac{-w(s) \left[\frac{(s+1)^2}{(s+2)^2} \right] \left[\frac{(s-2)}{(s+2)} \right] + \left[\frac{16}{3} \right]}{w(s) \left[\frac{(s+1)^2}{(s+2)^2} \right] \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right]} \end{aligned}$$

FEEDBACK SYSTEM DESIGN

$$= \frac{-w'(s) \left[\frac{(s-2)}{(s+2)} \right] + \left[\frac{16}{3} \right]}{w'(s) \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right]} \quad (2.39)$$

where

$$w'(s) = w(s) \left[\frac{(s+1)^2}{(s+2)^2} \right] \quad (2.40)$$

As such, the same set of compensators are obtained from the alternative coprime fractional representation as from the original representation though the parameterizations defined differ by the miniphase factor $[(s+1)^2]/[(s+2)^2]$.

Finally, let us set $w'(s) = 0$ in 2.39 obtaining the compensator

$$c(s) = \frac{16(s+2)}{3(s+2/3)} \quad (2.41)$$

Now, $c(s)$ has a zero at $s = -2$ which cancels the pole of $p(s)$ at $s = -2$. This does not, however, contradict the requirement that $p(s)c(s) = [n_p(s)n_c(s)]/[d_p(s)d_c(s)]$ be coprime since our coprimeness concept only forbids right half-plane pole-zero cancellations. Of course, a left half-plane pole-zero cancellation such as encountered in the above example is benign and need not be forbidden.

3. Tracking

Once we have stabilized our feedback system we may use the remaining design latitude, the choice of a stable $w(s)$, to meet various system design constraints. The first of several such design constraints which we will consider are the asymptotic tracking and disturbance rejection conditions wherein we require that the system asymptotically follow or reject a prescribed input [4].

In the *tracking* (or *asymptotic regulator*) problem it is desired to design a stable feedback system whose output, v_2 , asymptotically follows a prescribed input which we model by the impulse response of a transfer function $t(s)$, as illustrated in Figure 2. As usual we assume that $t(s)$ admits a

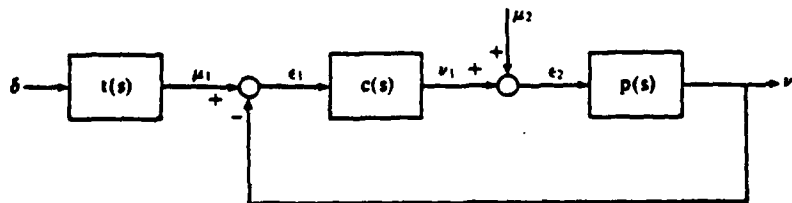


Figure 2. Feedback system with tracking generator.

coprime fractional representation in the form

$$t(s) = \frac{n_i(s)}{d_i(s)} \quad (3.1)$$

where there exist stable $u_i(s)$ and $v_i(s)$ such that

$$u_i(s)n_i(s) + v_i(s)d_i(s) = 1 \quad (3.2)$$

We say that the system tracks the impulse response of $t(s)$ if

$$t(s) - h_{n_i n_1}(s)t(s) = h_{v_i v_1}(s)t(s) \quad (3.3)$$

is stable. Recall that the impulse response of a single-variate system is asymptotic to zero if and only if the corresponding transfer function is stable. Thus the response of our system to the impulse response of $t(s)$ will be asymptotic to the impulse response of $t(s)$ if and only if the transfer of equation 3.2 is stable.

Recall from Corollary 6 that

$$h_{v_i v_1}(s) = w(s)n_p(s)d_p(s) + v_p(s)d_p(s) \quad (3.4)$$

hence if we desire to stabilize the system and simultaneously cause it to track the impulse response of $t(s)$ we must choose a stable $w(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero and

$$h_{n_i n_1}(s)t(s) = \frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)]\tilde{n}_i(s)}{d_i(s)} \quad (3.5)$$

is stable.

11. Property. Given $p(s)$ there exists a compensator for the feedback system of Figure 2 which stabilizes the system and simultaneously causes it to track the impulse response of $t(s)$ if and only if the equation

$$w(s)n_p(s)d_p(s) + x(s)d_i(s) = u_p(s)n_p(s) - 1$$

admits stable solutions $w(s)$ and $x(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. In this case the required compensator takes the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where $w(s)$ is a solution to the above equation.

Proof. If there exists a stable $w(s)$ such that 3.5 is stable then it follows from 3.2 that

$$\frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)]}{d_i(s)}$$

$$\begin{aligned}
 &= \frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)] [u_i(s)n_i(s) + v_i(s)d_i(s)]}{d_i(s)} \quad (3.6) \\
 &= \left[\frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)] n_i(s)}{d_i(s)} \right] u_i(s) + [w(s)n_p(s)d_p(s) + v_p(s)d_p(s)] v_i(s) \\
 &= -x(s)
 \end{aligned}$$

is stable since it is expressed as the sum of products of stable functions. Rearranging 3.6 and invoking 2.15 then yields

$$w(s)n_p(s)d_p(s) + x(s)d_i(s) = -v_p(s)d_p(s) = u_p(s)n_p(s) - 1 \quad (3.7)$$

as required. Conversely, if 3.7 admits stable solutions, $w(s)$ and $x(s)$, where $w(s)n_p(s) + v_p(s)$ is not identically zero we define $c(s)$ by

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]} \quad (3.8)$$

using the $w(s)$ of 3.7. Now, with this $w(s)$ 3.7 and 3.5 imply that

$$\begin{aligned}
 h_{e_1 n_i}(s) t(s) &= \frac{[w(s)n_p(s)d_p(s) + v_p(s)d_p(s)] n_i(s)}{d_i(s)} \quad (3.9) \\
 &= \frac{[-x(s)d_i(s)] n_i(s)}{d_i(s)} = -x(s)n_i(s)
 \end{aligned}$$

is stable. Since the stabilization theorem implies that any $c(s)$ in the form of 3.8 stabilizes the system while 3.9 implies that $h_{e_1 n_i}(s) t(s)$ is stable for this choice of $w(s)$ we have constructed the desired compensator.

Tracking Theorem: Given $p(s)$ there exists a compensator for the feedback system of Figure 2 which stabilizes the system and simultaneously causes it to track the impulse response of $t(s)$ if and only if $n_p(s)$ and $d_i(s)$ are coprime. In this case let $u_{pi}(s)$ and $v_{pi}(s)$ be stable functions such that

$$u_{pi}(s)n_p(s) + v_{pi}(s)d_i(s) = 1$$

and let $a(s) = n_o(s)/d_o(s)$ be a coprime fractional representation of $a(s) = d_p(s)/d_i(s)$. Then the desired set of compensators take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where

$$w(s) = -u_{pi}(s)v_p(s) + e(s)d_o(s)$$

with $e(s)$ an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero.

Proof. Consistent with Property 11, it suffices to show that the equation

$$w(s)n_p(s)d_p(s) + x(s)d_i(s) = u_p(s)n_p(s) - 1 \quad (3.10)$$

admits stable solutions $w(s)$ and $x(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero if and only if $n_p(s)$ and $d_i(s)$ are coprime. In this case we then show that the appropriate $w(s)$ takes the form

$$w(s) = -u_{pr}(s)v_p(s) + e(s)d_p(s) \quad (3.11)$$

for any stable $e(s)$.

If 3.10 admits stable solutions $w(s)$ and $x(s)$ then it follows from 3.10 that

$$u_p(s)n_p(s) - w(s)n_p(s)d_p(s) - x(s)d_i(s) = 1 \quad (3.12)$$

or equivalently

$$[u_p(s) - w(s)d_p(s)]n_p(s) + [-x(s)]d_i(s) = 1 \quad (3.13)$$

showing that $n_p(s)$ and $d_i(s)$ are coprime. Conversely, if $n_p(s)$ and $d_i(s)$ are coprime there exists $u_{pr}(s)$ and $v_{pr}(s)$ such that

$$u_{pr}(s)n_p(s) + v_{pr}(s)d_i(s) = 1 \quad (3.14)$$

from which it follows that

$$\begin{aligned} u_p(s)n_p(s) - 1 &= -v_p(s)d_p(s) \\ &= -[u_{pr}(s)n_p(s) + v_{pr}(s)d_i(s)v_p(s)]v_p(s)d_p(s) \\ &= [-u_{pr}(s)v_p(s)]n_p(s)d_p(s) + [-v_{pr}(s)v_p(s)d_p(s)]d_i(s) \end{aligned} \quad (3.15)$$

As such,

$$w^p(s) = -u_{pr}(s)v_p(s) \quad (3.16)$$

and

$$x^p(s) = -v_{pr}(s)v_p(s)d_p(s) \quad (3.17)$$

represent particular solutions to 3.10.

To construct homogeneous solutions 3.10 we define a transfer function $a(s)$ by $a(s) = d_p(s)/d_i(s)$ and let

$$a(s) = \frac{n_p(s)}{d_i(s)} \quad (3.18)$$

be a coprime fractional representation for $a(s)$. It then follows from Property 2 that there exists a stable $k(s)$ such that

$$d_p(s) = n_a(s)k(s) \quad (3.19)$$

and

$$d_i(s) = d_a(s)k(s) \quad (3.20)$$

Thus if we define candidates for the homogeneous solution of 3.10 by

$$w^h(s) = e(s)d_a(s) \quad (3.21)$$

and

$$x^h(s) = -e(s)n_p(s)n_a(s) \quad (3.22)$$

where $e(s)$ is an arbitrary stable rational function we have

$$w^h(s)n_p(s)d_p(s) + x^h(s)d_i(s) = 0$$

$$= e(s)d_a(s)n_p(s)n_a(s)k(s) - e(s)n_p(s)n_a(s)d_a(s)k(s) = 0$$

verifying that 3.21 and 3.22 are, indeed, homogeneous solutions.

To complete the solution of 3.10 we must show that all homogeneous solutions are of the form 3.21 and 3.22. To this end assume that $\underline{w}^h(s)$ and $\underline{x}^h(s)$ are stable and satisfy

$$\underline{w}^h(s)n_p(s)d_p(s) + \underline{x}^h(s)d_i(s) = 0 \quad (3.24)$$

and define $e(s)$ by

$$e(s) = \underline{w}^h(s)d_a(s) \quad (3.25)$$

Clearly,

$$\underline{w}^h(s) = e(s)d_a(s) \quad (3.26)$$

while it follows from 3.24 that

$$\begin{aligned} \underline{x}^h(s) &= \frac{-\underline{w}^h(s)n_p(s)d_p(s)}{d_i(s)} = -\underline{w}^h(s)n_p(s) \\ &= \frac{-\underline{w}^h(s)n_p(s)n_a(s)}{d_a(s)} = -e(s)n_p(s)n_a(s) \end{aligned} \quad (3.27)$$

showing that $w^h(s)$ and $x^h(s)$ have the form of 3.21 and 3.22. It remains, however, to verify that $e(s)$ is stable for which purpose we invoke equations 3.14 and the coprimeness of $n_a(s)$ and $d_a(s)$ from which it follows that there exists stable $u_a(s)$ and $v_a(s)$ such that

$$u_a(s)n_a(s) + v_a(s)d_a(s) = 1 \quad (3.28)$$

As such,

$$\begin{aligned}
 e(s) &= \frac{w^h(s)}{d_e(s)} = \left[\frac{w^h(s)}{d_e(s)} \right] [u_e(s)n_e(s) + v_e(s)d_e(s)] \\
 &= \frac{w^h(s)u_e(s)n_e(s)}{d_e(s)} + w^h(s)v_e(s) \\
 &= w^h(s)u_e(s)a(s) + w^h(s)v_e(s) = \frac{w^h(s)u_e(s)d_p(s)}{d_i(s)} + w^h(s)v_e(s) \quad (3.29) \\
 &= \left[\frac{w^h(s)u_e(s)d_p(s)}{d_i(s)} \right] [u_{pi}(s)n_p(s) + v_{pi}(s)d_i(s)] + w^h(s)v_e(s) \\
 &= \frac{w^h(s)u_e(s)d_p(s)u_{pi}(s)n_p(s)}{d_i(s)} + w^h(s)u_e(s)d_p(s)v_{pi}(s) + w^h(s)v_e(s) \\
 &= -x^h(s)u_e(s)u_{pi}(s) + w^h(s)u_e(s)d_p(s)v_{pi}(s) + w^h(s)v_e(s)
 \end{aligned}$$

is stable since we have expressed it as the sum of products of stable functions. Note, the last equality in 3.29 follows from 3.24.

The solution space for Equation 3.10 thus takes the form

$$w(s) = -u_{pi}(s)v_p(s) + e(s)d_e(s) \quad (3.30)$$

and

$$x(s) = -v_{pi}(s)v_p(s)d_p(s) - e(s)n_p(s)n_e(s) \quad (3.31)$$

As such, Property 11 implies that if the $w(s)$ of Equation 3.30 is used to define a stabilizing compensator as per the stabilization theorem it will also cause the system to track the impulse response of $T(s)$. Of course, we must also assume that $e(s)$ is chosen so that $w(s)n_p(s) + v_p(s)$ is not identically zero. To complete our proof that the coprimeness of $n_p(s)$ and $d_i(s)$ is a sufficient condition for the solution of the tracking problem it thus suffices to show that there exists at least one choice of $e(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. From 3.30 it follows that

$$w(s)n_p(s) + v_p(s) = [-u_{pi}(s)n_p(s) + 1]v_p(s) + e(s)d_e(s)n_p(s) \quad (3.32)$$

Now, $d_e(s)$ is not identically zero since it is the denominator of $a(s)$. As such, if $n_p(s)$ is not identically zero it follows from 3.32 that the value of $w(s)n_p(s) + v_p(s)$ is a non-trivial function of $e(s)$ and is therefore not identical to zero for all $e(s)$. On the other hand if $n_p(s) = 0$ then 2.15 implies that $v_p(s)$ is miniphase which, in turn, implies that $w(s)n_p(s) + v_p(s) = v_p(s)$ is not identically zero. Our proof is therefore complete.

Although the proof of our theorem is long, though elementary, the basic

result to the effect that a compensator exists which will simultaneously stabilize the system and cause it to track the impulse of $t(s)$ if and only if $n_p(s)$ and $d_p(s)$ are coprime is simple to check (no common right half-plane zeros). Moreover, the construction of the required compensator is simply a matter of substitution as per the following example.

12. Example. Consider the problem of designing a compensator for the plant of Examples 4 and 10 so that the system is stable and asymptotically tracks a step function. Recall that

$$p(s) = \frac{\begin{bmatrix} (s+1) \\ (s^2-4) \end{bmatrix}}{\begin{bmatrix} (s-2) \\ (s+2) \end{bmatrix}} = \frac{n_p(s)}{d_p(s)} \quad (3.33)$$

where

$$\frac{16}{3} \frac{\begin{bmatrix} (s+1) \\ (s+2)^2 \end{bmatrix}}{\begin{bmatrix} (s+2/3) \\ (s+2) \end{bmatrix}} + \frac{\begin{bmatrix} (s-2) \\ (s+2) \end{bmatrix}}{\begin{bmatrix} (s-2) \\ (s+2) \end{bmatrix}} = u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (3.34)$$

While we take $t(s) = 1/s$ from which we obtain

$$t(s) = \frac{1}{s} = \frac{\begin{bmatrix} 1 \\ (s+2) \end{bmatrix}}{\begin{bmatrix} s \\ (s+2) \end{bmatrix}} = \frac{n_t(s)}{d_t(s)} \quad (3.35)$$

Now, $n_p(s)$ has no right half-plane zeros while $d_t(s)$ has a right half-plane zero at $s = 0$. As such, $n_p(s)$ and $d_t(s)$ are coprime and according to the theorem the desired compensator exists. Indeed,

$$\begin{bmatrix} 4 \end{bmatrix} \frac{\begin{bmatrix} (s+1) \\ (s+2)^2 \end{bmatrix}}{\begin{bmatrix} (s+2/3) \\ (s+2) \end{bmatrix}} + \begin{bmatrix} s \\ (s+2) \end{bmatrix} \frac{\begin{bmatrix} s \\ (s+2) \end{bmatrix}}{\begin{bmatrix} s \\ (s+2) \end{bmatrix}} = u_{pt}(s)n_p(s) + v_{pt}(s)d_t(s) = 1 \quad (3.36)$$

As a final step in the construction of our compensator we let

$$a(s) = \frac{d_p(s)}{d_t(s)} = \frac{\begin{bmatrix} (s-2) \\ (s+2) \end{bmatrix}}{\begin{bmatrix} s \\ (s+2) \end{bmatrix}} = \frac{n_a(s)}{d_a(s)} \quad (3.37)$$

which is coprime since

$$\begin{bmatrix} -1 \end{bmatrix} \frac{\begin{bmatrix} (s-2) \\ (s+2) \end{bmatrix}}{\begin{bmatrix} (s+2) \end{bmatrix}} + \begin{bmatrix} 2 \end{bmatrix} \frac{\begin{bmatrix} 2 \\ (s+2) \end{bmatrix}}{\begin{bmatrix} (s+2) \end{bmatrix}} = u_a(s)n_a(s) + v_a(s)d_a(s) \quad (3.38)$$

It thus follows from the theorem that the desired $w(s)$ takes the form

$$w(s) = -u_p(s)v_p(s) + e(s)d_g(s) = \left[\frac{-4(s+2/3)}{(s+2)} \right] + \left[\frac{s}{(s+2)} \right] e(s) \quad (3.39)$$

where $e(s)$ is an arbitrary stable function. Substitution of any such $w(s)$ into

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]} \quad (3.40)$$

thus defines the required compensator so long as $e(s)$ is chosen so that $w(s)n_p(s) + v_p(s)$ is not identically zero. To verify our solution we may substitute $w(s)$ into the formula for $h_{1,\mu_1}(s)$ of Corollary 6 obtaining

$$h_{1,\mu_1}(s) = \left[\frac{s}{(s+2)^4} \right] [s(s-2)(s+2/3) + (s+1)(s-2)e(s)] \quad (3.41)$$

Since $h_{1,\mu_1}(s)$ has a zero at $s = 0$ the required tracking property may now be verified by the final value theorem.

Now let us consider an alternative problem where we are required to track $e^{2t} \cup(t)$. Here our tracking generator is defined by

$$r(s) = \left[\frac{1}{(s-2)} \right] = \frac{\left[\frac{1}{(s+2)} \right]}{\left[\frac{(s-2)}{(s+2)} \right]} = \frac{\bar{n}_r(s)}{d_r(s)} \quad (3.42)$$

As before $n_p(s)$ and $d_r(s)$ are coprime since

$$\left[\frac{16}{3} \right] \left[\frac{(s+1)}{(s+2)^2} \right] + \left[\frac{(s+2/3)}{(s+2)} \right] \left[\frac{(s-2)}{(s+2)} \right] = u_p(s)n_p(s) + v_p(s)d_r(s) = 1 \quad (3.43)$$

Finally, for this example we have

$$a(s) = \frac{d_p(s)}{d_r(s)} = \frac{\left[\frac{(s-2)}{(s+2)} \right]}{\left[\frac{(s-2)}{(s+2)} \right]} = \frac{1}{1} = \frac{n_g(s)}{d_g(s)} \quad (3.44)$$

where

$$[0][1] + [1][1] = u_g(s)n_g(s) + v_g(s)d_g(s) = 1 \quad (3.45)$$

Note that in this example $d_p(s)$ and $d_r(s)$ are not coprime since they have a common right half-plane zero at $s = 2$. For our purposes, however, all that is required is a coprime fractional representation for $a(s) = d_p(s)/d_r(s)$ as constructed above. Using these new values for $u_p(s)$ and $d_g(s)$ we obtain

$$w(s) = - \left[\frac{16(s+2/3)}{3(s+2)} \right] + e(s) \quad (3.46)$$

for any stable $e(s)$. This, in turn, yields

$$h_{e_1 \mu_1}(s) = \left[\frac{(s-2)}{9(s+2)^4} \right] [(9s^3 - 6s^2 - 20s - 8) + 9(s+1)(s+2)e(s)] \quad (3.47)$$

for which the zero at $s = 2$ indicates tracking. Note that every stable $w(s)$ is obtained for some $e(s)$ and hence all stabilizing compensators track $e^{2t} \cup (t)$ in this example.

4. Disturbance rejection

There are two alternative disturbance rejection problems which arise naturally in our feedback system theory. Figure 3a indicates the configuration for the *input disturbance rejection problem* [16] wherein we desire to design a compensator which simultaneously stabilizes the system and causes it to asymptotically reject the impulse response of $r(s)$, i.e., the response of the system to the impulse response of $r(s)$ should be asymptotic to zero.

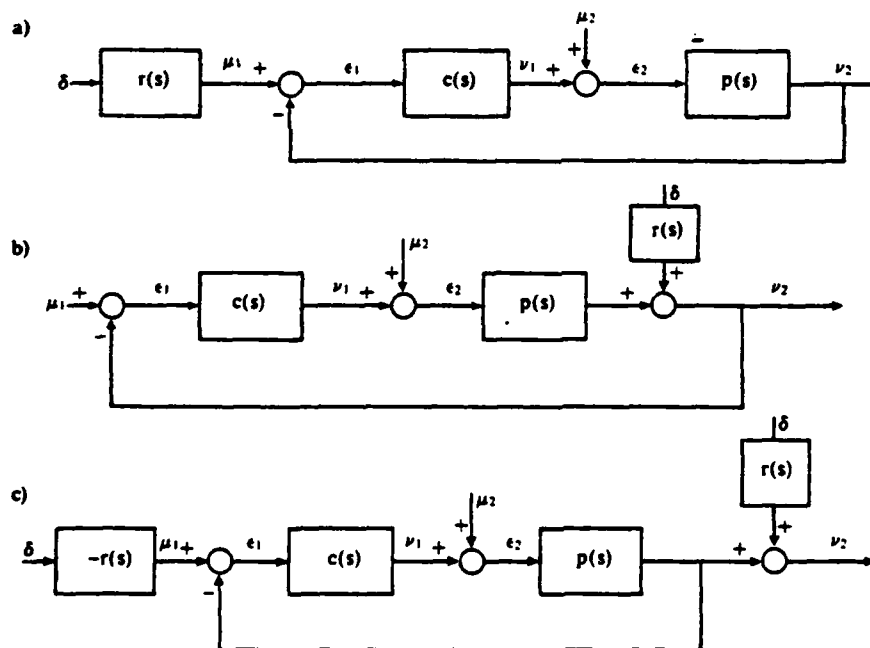


Figure 3. Feedback system configuration for a) the input disturbance rejection problem, b) the output disturbance rejection problem, and c) a modified configuration for the output disturbance rejection problem.

A similar *output disturbance rejection problem* [4] is illustrated in Figure 3b. Here, the disturbance is injected into the system at the plant output and, as before, it is desired to design a compensator which simultaneously stabilizes the system and causes it to asymptotically reject the impulse response of $r(s)$. Surprisingly, however, the output disturbance rejection problem is completely equivalent to the tracking problem considered in the previous section. To see this simply observe that the block diagram of Figure 3c is equivalent to that of Figure 3b. As such, if we design a compensator to stabilize the system and cause the plant output to asymptotically track the impulse response of $-r(s)$ when the impulse response of $r(s)$ is added to the plant output the total effect of the disturbance observed at y_2 will be asymptotic to zero. Consistent with the above we give no further consideration to the output disturbance rejection problem since it may be resolved via the techniques of the previous section with $t(s) = -r(s)$. Indeed, one can solve the tracking and output rejection problem simultaneously by working with tracking generator $t(s) = r(s)$.

For the input disturbance rejection problem we require that the impulse response of

$$h_{y_2, u_1}(s)r(s) = [-w(s)n_p(s)d_p(s) + u_p(s)n_p(s)]r(s) \quad (4.1)$$

be asymptotic to zero. Hence to simultaneously stabilize the feedback system and cause it to asymptotically reject the impulse response of $r(s)$ we must choose a stable $w(s)$ such that $w(s)n_p(s) + u_p(s)$ is not identically zero and $h_{y_2, u_1}(s)r(s)$ is stable. The required theory [10] is essentially identical to that used to solve the tracking problem and hence we simply state the pertinent theorems without proof. For this we let $r(s) = n_r(s)/d_r(s)$ be a coprime fractional representation for $r(s)$.

13. Property. Given $p(s)$ there exists a compensator for the feedback system of Figure 3a which stabilizes the system and simultaneously causes it to reject the impulse response of $r(s)$ if and only if the equation

$$w(s)n_p(s)d_p(s) + y(s)d_r(s) = u_p(s)n_p(s)$$

admits stable solutions $w(s)$ and $y(s)$ such that $w(s)n_p(s) + u_p(s)$ is not identically zero. In this case the required compensator takes the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + u_p(s)]}$$

where $w(s)$ is a solution to the above equation.

Disturbance Rejection Theorem. Given $p(s)$ there exists a compensator for the feedback system of Figure 3a which stabilizes the system and simultaneously causes it to reject the impulse response of $r(s)$ if and only if $d_p(s)$ and $d_r(s)$ are coprime. In this case let $u_{pr}(s)$ and $v_{pr}(s)$

be stable functions such that

$$u_{pr}(s)d_p(s) + v_{pr}(s)d_r(s) = 1$$

and let $b(s) = n_b(s)/d_b(s)$ be a coprime fractional representation of $b(s) = n_p(s)/d_r(s)$. Then the desired set of compensators take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where

$$w(s) = u_{pr}(s)u_p(s) + f(s)d_b(s)$$

with $f(s)$ an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero.

14. Example. Continuing with the plant of Example 12 let us consider the problem of designing a compensator to reject a step function, i.e., we let

$$r(s) = \begin{bmatrix} 1 \\ s \end{bmatrix} = \frac{\begin{bmatrix} 1 \\ (s+2) \end{bmatrix}}{\begin{bmatrix} s \\ (s+2) \end{bmatrix}} = \frac{n_r(s)}{d_r(s)}$$

Now,

$$\begin{bmatrix} -1 \\ 2 \end{bmatrix} \begin{bmatrix} (s-2) \\ (s+2) \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} s \\ (s+2) \end{bmatrix} = u_{pr}(s)d_p(s) + v_{pr}(s)d_r(s) = 1 \quad (4.3)$$

showing that $d_p(s)$ and $d_r(s)$ are coprime. Finally, we let

$$b(s) = \frac{n_p(s)}{d_r(s)} = \frac{\begin{bmatrix} (s+1) \\ (s+2)^2 \end{bmatrix}}{\begin{bmatrix} s \\ (s+2) \end{bmatrix}} = \frac{n_b(s)}{d_b(s)} \quad (4.4)$$

which is clearly coprime. From the theorem the required $w(s)$ take the form

$$w(s) = u_{pr}(s)u_p(s) + f(s)d_b(s) = \begin{bmatrix} -16 \\ 3 \end{bmatrix} + \begin{bmatrix} s \\ (s+2) \end{bmatrix} f(s) \quad (4.5)$$

where $f(s)$ is arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero. The use of the compensator derived from this $w(s)$ then results in the gains

$$h_{r2\mu_1}(s) = \begin{bmatrix} s(s+1) \\ (s+2)^4 \end{bmatrix} \begin{bmatrix} 32 \\ 3 \end{bmatrix} (s+2) - (s-2)f(s) \quad (4.6)$$

and

$$h_{1,\mu_1}(s) = \left[\frac{(s-2)}{3(s+2)^4} \right] [(s+2)(3s^2 - 8s - 12) + 3s(s+1)f(s)] \quad (4.7)$$

Here, the fact that $h_{2,\mu_1}(s)$ has a zero at $s = 0$ indicates that the disturbance rejection specification has been achieved while the fact that $h_{1,\mu_1}(s)$ has a zero at $s = 2$ implies that the system also tracks $e^{2t} \cup (t)$. This is consistent with Example 12 for the system tracks $e^{2t} \cup (t)$.

5. Simultaneous tracking and disturbance rejection

The purpose of this section is to combine the results of the previous two sections by formulating criteria for the design of a compensator which simultaneously stabilizes the system, causes it to track the impulse response of $r(s)$ and causes it to reject the impulse response of $r(s)$ [16]. The appropriate feedback system configuration is shown in Figure 4 where $r(s)$ is taken to be an input disturbance. Of course, an output disturbance can also be included in the theory simply by combining it with the tracking generator.

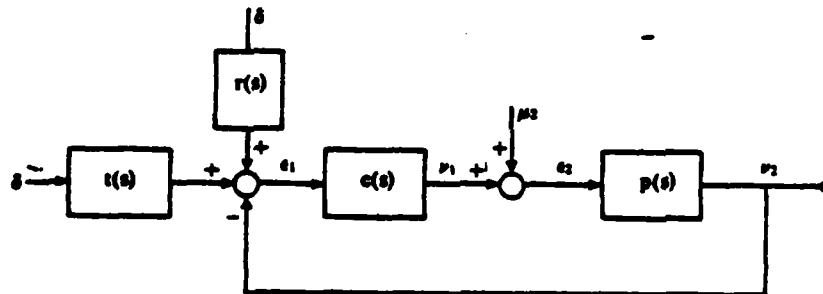


Figure 4. Configuration for the simultaneous tracking and disturbance rejection problem.

For consistency with the previous sections we will use the same notation which is reviewed as follows. Our plant is assumed to have a coprime fractional representation

$$p(s) = \frac{n_p(s)}{d_p(s)} \quad (5.1)$$

such that

$$u_p(s)n_p(s) + v_p(s)d_p(s) = 1 \quad (5.2)$$

while the tracking and rejection generators are characterized by coprime fractional representations

$$r(s) = \frac{n_r(s)}{d_r(s)} \quad (5.3)$$

and

$$r(s) = \frac{n_r(s)}{d_r(s)} \quad (5.4)$$

Also as in the previous sections we define $a(s)$ and $b(s)$ by

$$a(s) = \frac{d_p(s)}{d_r(s)} = \frac{n_a(s)}{d_a(s)} \quad (5.5)$$

and

$$b(s) = \frac{d_p(s)}{d_r(s)} = \frac{n_b(s)}{d_b(s)} \quad (5.6)$$

where $a(s) = n_a(s)/d_a(s)$ and $b(s) = n_b(s)/d_b(s)$ are coprime in the sense that there exists stable $u_a(s)$, $v_a(s)$, $u_b(s)$, and $v_b(s)$ such that

$$u_a(s)n_a(s) + v_a(s)d_a(s) = 1 \quad (5.7)$$

and

$$u_b(s)n_b(s) + v_b(s)d_b(s) = 1 \quad (5.8)$$

Moreover it follows from 5.5 and 5.6 together with property 2 that there exists stable $k(s)$ and $m(s)$ such that

$$d_r(s) = d_a(s)k(s) \text{ and } d_p(s) = n_a(s)k(s) \quad (5.9)$$

and

$$d_r(s) = d_b(s)m(s) \text{ and } n_p(s) = n_b(s)m(s) \quad (5.10)$$

Finally, the coprimeness conditions for the tracking and disturbance rejection problems are characterized by

$$u_{pr}(s)n_p(s) + v_{pr}(s)d_r(s) = 1 \quad (5.11)$$

and

$$u_{pr}(s)d_p(s) + v_{pr}(s)d_r(s) = 1 \quad (5.12)$$

while we will also require a coprimeness condition between $d_r(s)$ and $d_p(s)$ which we characterize by the equation

$$u_{tr}(s)d_r(s) + v_{tr}(s)d_p(s) = 1 \quad (5.13)$$

With this review of notation in hand the required design equations for the simultaneous tracking and disturbance rejection problem can be obtained simply by combining the results of Property 11 and Property 13 and observing that both design equations must be satisfied by the same $w(s)$ since we desire to construct a single compensator which simultaneously solves both problems.

15. Property. Given $p(s)$ there exists a compensator for the feedback system of Figure 4 which stabilizes the system, causes it to track the impulse response of $r(s)$ and simultaneously causes it to reject the impulse response of $r(s)$ if and only if the pair of equations

$$w(s)n_p(s)d_p(s) + x(s)d_r(s) = u_p(s)n_p(s) - 1$$

and

$$w(s)n_p(s)d_p(s) + y(s)d_r(s) = u_p(s)n_p(s)$$

admit stable solutions $w(s)$, $x(s)$, and $y(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. In this case the required compensator takes the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where $w(s)$ is a solution to the above equations.

Simultaneous Tracking and Disturbance Rejection Theorem: Given $p(s)$ there exists a compensator for the feedback system of Figure 4 which stabilizes the system, causes it to track the impulse response of $r(s)$ and simultaneously causes it to reject the impulse response of $r(s)$ if and only if

- (i) $n_p(s)$ and $d_r(s)$ are coprime,
- (ii) $d_p(s)$ and $d_r(s)$ are coprime, and
- (iii) $d_p(s)$ and $d_r(s)$ are coprime.

In that case the desired set of compensators take the form

$$c(s) = \frac{[-w(s)d_p(s) + u_p(s)]}{[w(s)n_p(s) + v_p(s)]}$$

where

$$w(s) = [u_{pr}(s)u_p(s)u_r(s)d_r(s) - u_{pr}(s)v_p(s)v_r(s)d_r(s)] + g(s)d_p(s)d_r(s)$$

with $g(s)$ an arbitrary stable function such that $w(s)n_p(s) + v_p(s)$ is not identically zero.

Proof. The fact that $n_p(s)$ and $d_r(s)$ must be coprime follows from the tracking theorem while the fact that $d_p(s)$ and $d_r(s)$ must be coprime follows from the disturbance rejection theorem. To verify that $d_p(s)$ and $d_r(s)$ must also be coprime for simultaneous stabilization we subtract the two design equations of Property 15 obtaining

$$[y(s)]d_r(s) + [-x(s)]d_r(s) = 1 \quad (5.14)$$

Conversely, to show that the three coprimeness conditions are also sufficient we must construct a $w(s)$ which simultaneously satisfies the criteria for tracking and disturbance rejections derived in the preceding sections. Upon invoking the results of the tracking and disturbance rejection

theorems we must therefore solve

$$-u_{pr}(s)v_p(s) + e(s)d_a(s) = w(s) = u_{pr}(s)u_p(s) + f(s)d_b(s) \quad (5.15)$$

for stable $e(s)$, $f(s)$, and $w(s)$. Since the required set of stable $w(s)$ may be obtained by substitution once $e(s)$ and $f(s)$ have been parameterized our main problem is to characterize the stable solutions of

$$e(s)d_a(s) - f(s)d_b(s) = \{u_{pr}(s)u_p(s) + u_{pi}(s)v_p(s)\} \quad (5.16)$$

To obtain a particular solution for 5.16 we invoke 5.9, 5.10, and 5.13 obtaining

$$\begin{aligned} & [u_{pr}(s)u_p(s) + u_{pi}(s)v_p(s)] \\ &= [u_{pr}(s)u_p(s) + u_{pi}(s)v_p(s)] [u_{ir}(s)d_i(s) + v_{ir}(s)d_r(s)] \\ &= [u_{pr}(s)u_p(s) + u_{pi}(s)v_p(s)] [u_{ir}(s)k(s)d_a(s) + v_{ir}(s)m(s)d_b(s)] \\ &= \{[u_{pr}(s)u_p(s) + u_{pi}(s)v_p(s)]u_{ir}(s)k(s)\}d_a(s) \\ & \quad + \{[u_{pr}(s)u_p(s) + u_{pi}(s)v_p(s)]v_{ir}(s)m(s)\}d_b(s) \end{aligned} \quad (5.17)$$

As such, the required particular solutions take the form

$$e^p(s) = \{u_{pr}(s)u_p(s) + u_{pi}(s)v_p(s)\}u_{ir}(s)k(s) \quad (5.18)$$

and

$$f^p(s) = -\{u_{pr}(s)u_p(s) + u_{pi}(s)v_p(s)\}v_{ir}(s)m(s) \quad (5.19)$$

Of course,

$$e^h(s) = g(s)d_b(s) \quad (5.20)$$

and

$$f^h(s) = g(s)d_a(s) \quad (5.21)$$

represent homogeneous solutions to 5.16 for any stable $g(s)$ since

$$e^h(s)d_a(s) - f^h(s)d_b(s) = g(s)d_b(s)d_a(s) - g(s)d_a(s)d_b(s) = 0 \quad (5.22)$$

As such, our characterization of the solution space for 5.16 will be complete if we can verify that all stable homogeneous solutions to 5.16 take the form of 5.20 and 5.21 for some stable $g(s)$. To this end let $\underline{e}^h(s)$ and $\underline{f}^h(s)$ be arbitrary stable homogeneous solutions to 5.16, i.e.,

$$\underline{e}^h(s)d_a(s) - \underline{f}^h(s)d_b(s) = 0 \quad (5.23)$$

Now, define $g(s)$ by $g(s) = \underline{e}^h(s)/d_b(s)$ in which case

$$\underline{e}^h(s) = g(s)d_b(s) \quad (5.24)$$

while 5.23 implies that

$$\underline{f}^h(s) = \frac{\underline{e}^h(s) d_a(s)}{d_b(s)} = g(s) d_a(s) \quad (5.25)$$

showing that $\underline{e}^h(s)$ and $\underline{f}^h(s)$ are of the required form. It remains, however, to show that $g(s)$ is stable. Indeed,

$$\begin{aligned} g(s) &= \frac{\underline{e}^h(s)}{d_b(s)} = \frac{\underline{e}^h(s)}{d_b(s)} [u_b(s) n_b(s) + v_b(s) d_b(s)] \\ &= \underline{e}^h(s) v_b(s) + \frac{\underline{e}^h(s) u_b(s) n_b(s)}{d_b(s)} \\ &= \underline{e}^h(s) v_b(s) + \frac{\underline{e}^h(s) u_b(s) n_p(s)}{d_r(s)} [u_{ir}(s) d_i(s) + v_{ir}(s) d_r(s)] \\ &= \underline{e}^h(s) v_b(s) + \underline{e}^h(s) u_b(s) n_p(s) v_{ir}(s) + \frac{\underline{e}^h(s) u_b(s) n_p(s) u_{ir}(s) d_i(s)}{d_r(s)} \\ &= \underline{e}^h(s) v_b(s) + \underline{e}^h(s) u_b(s) n_p(s) v_{ir}(s) + \frac{\underline{e}^h(s) u_b(s) n_b(s) u_{ir}(s) d_a(s) k(s)}{d_b(s)} \\ &= \underline{e}^h(s) v_b(s) + \underline{e}^h(s) u_b(s) n_p(s) v_{ir}(s) + \underline{f}^h(s) u_b(s) n_b(s) u_{ir}(s) k(s) \quad (5.26) \end{aligned}$$

showing that $g(s)$ is stable since it has been expressed as the sum of products of stable functions. Here, Equation 5.26 was derived with the aid of Equations 5.6, 5.8, 5.9, 5.13, and 5.23. As such, the complete set of solutions to 5.16 take the form

$$e(s) = [u_{pr}(s) u_p(s) + u_{pr}(s) v_p(s)] u_{ir}(s) k(s) + g(s) d_b(s) \quad (5.27)$$

and

$$f(s) = [u_{pr}(s) u_p(s) + u_{pr}(s) v_p(s)] v_{ir}(s) m(s) + g(s) d_a(s) \quad (5.28)$$

Now, upon substituting either of these expressions into 5.15 we obtain the desired expression for $w(s)$ in the form

$$\begin{aligned} w(s) &= -u_{pr}(s) v_p(s) + u_{pr}(s) u_p(s) + u_{pr}(s) v_p(s) u_{ir}(s) k(s) d_a(s) + g(s) d_b(s) d_a(s) \\ &= u_{pr}(s) u_p(s) u_{ir}(s) d_i(s) + u_{pr}(s) v_p(s) u_{ir}(s) d_i(s) - 1 + g(s) d_b(s) d_a(s) \\ &= [u_{pr}(s) u_p(s) u_{ir}(s) d_i(s) - u_{pr}(s) v_p(s) v_{ir}(s) d_r(s)] + g(s) d_b(s) d_a(s) \quad (5.29) \end{aligned}$$

where $g(s)$ is an arbitrary function.

Finally, to complete the proof that the three coprimeness conditions suffice for simultaneous tracking and disturbance rejection we must verify that there exists one choice of a stable $g(s)$ such that $w(s)n_p(s) + v_p(s)$ is not identically zero. Upon substituting 5.29 into this expression we obtain

$$\begin{aligned} w(s)n_p(s) + v_p(s) = & [u_{pr}(s)u_p(s)u_{ir}(s)d_i(s) - u_{pr}(s)v_p(s)v_{ir}(s)d_r(s)]n_p(s) \\ & + v_p(s) + g(s)d_b(s)d_e(s)n_p(s) \end{aligned} \quad (5.30)$$

Now $d_b(s)$ and $d_e(s)$ are not identically zero since they represent the denominators for well defined transfer functions. As such, if $n_p(s)$ is not identically zero 5.30 will be non-trivially dependent on $g(s)$ and hence not identically zero for every choice of $g(s)$. On the other hand if $n_p(s)$ is identically zero the equality $u_p(s)n_p(s) + v_p(s)d_p(s) = 1$ implies that $v_p(s)$ is miniphase hence

$$w(s)n_p(s) + v_p(s) = v_p(s) \quad (5.31)$$

is not identically zero. Our proof is therefore complete.

Although the theorem is highly complex and, indeed, is predicated on the equally complex theorems which preceded it, the final result is an explicit description of the desired family of compensators. Moreover, the terms in this expression are readily computed by solving one or more coprimeness equation. As such, the result is easily implemented as per the following example.

16. Example. Continuing our analysis of the plant introduced in the previous examples we will investigate the possibility of simultaneously tracking $e^{2t}U(t)$ and rejecting $U(t)$. Here,

$$d_i(s) = \frac{(s-2)}{(s+2)} \quad (5.32)$$

and

$$d_r(s) = \frac{s}{(s+2)} \quad (5.33)$$

which are clearly coprime. Indeed

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix} \begin{bmatrix} (s-2) \\ (s+2) \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} s \\ (s+2) \end{bmatrix} = u_{ir}(s)d_i(s) + v_{ir}(s)d_r(s) = 1 \quad (5.34)$$

As such, the required $w(s)$ takes the form

$$w(s) = \left[\frac{1}{(s+2)^2} \right] \left[\frac{-16}{3}(s^2 + \frac{4}{3}s + 4) + s(s+2)g(s) \right] \quad (5.35)$$

yielding

$$h_{2\mu_1}(s) = \left[\frac{(s+1)s}{(s+2)^5} \right] \left[\frac{32}{3} \left(s^2 + \frac{8}{3}s + \frac{20}{3} \right) - (s-2)(s+2)g(s) \right] \quad (5.36)$$

and

$$h_{1\mu_1}(s) = \left[\frac{(s-2)}{9(s+2)^5} \right] \left[(9s^4 + 12s^3 + 32s^2 - 112s - 144) + 9s(s+1)(s+2)g(s) \right] \quad (5.37)$$

which have the required zeros at $s = 0$ and $s = 2$, respectively.

References

1. Antsaklis, P.J., and J.B. Pearson, "Stabilization and Regulation in Linear Multivariate Systems", IEEE Trans. on Auto. Cont., Vol. AC-23, pp. 928-930, (1978).
2. Barnett, S. *Matrices in Control Theory*, London, Van Nostrand, 1971.
3. Bengtsson, G., "Output Regulation and Internal Modes - A Frequency Domain Approach", Automatica, Vol. 13, pp. 333-345, (1977).
4. Callier, F.M., and C.A. Desoer, "Stabilization, Tracking and Disturbance Rejection in Linear Multivariable Distributed Systems" Proc. of the 17th IEEE Conf. on Decision and Control, San Diego, Jan. 1979, pp. 513-514, (also Tech Memo UCB/ERL M78/83, Univ. of California at Berkeley, Dec. 1978).
5. Chang, L., and J.B. Pearson, "Frequency Domain Synthesis of Multivariable Linear Regulators", IEEE Trans. on Auto. Cont., Vol. AC-23, pp. 3-15, (1978).
6. Cheng, L., and J.B. Pearson, "Synthesis of Linear Multivariable Regulators", IEEE Trans. on Auto. Cont., (to appear).
7. Chua, O., M.S. Thesis, Texas Tech Univ., 1980.
8. Desoer, C.A., Liu, R.-w., Murray, J., and R. Saecks, "Feedback System Design: the fractional representation approach to analysis and synthesis", IEEE Trans. on Auto. Cont., Vol. AC-25, pp. 399-412, (1980).
9. Feintuch, A., and R. Saecks, *System Theory: A Hilbert Space Approach*, New York, Academic Press, (to appear).
10. Francis, B., "The Multivariable Servomechanism Problem from the Input-Output Viewpoint", IEEE Trans. on Auto. Cont., Vol. AC-22, pp. 322-328, (1977).
11. Francis, B., and M. Vidyasagar, "Algebraic and Topological Aspects of the Servo Problem for Lumped Linear Systems", unpublished notes, Yale Univ., 1980.
12. Helton, J.W., "Orbit Structure of the Moebius Transformation Semigroup Acting on H^{∞} " in *Topics in Functional Analysis*, Advances in Mat. Supp. Studies, Vol. 3, New York, Academic Press, 1978, pp. 129-157.
13. Pernebo, L., Ph.D. Thesis, Lund Inst. of Tech., 1978.
14. Pernebo, L., "An Algebraic Theory for Design of Controllers for Linear Multivariable Systems; Parts I and II", IEEE Trans. on Auto. Cont., (to appear).

FEEDBACK SYSTEM DESIGN

15. Rosenbrock, H.H., *State-Space and Multivariable Theory*, New York, J. Wiley and Sons, 1970.
16. Sacks, R., and J. Murray, "Feedback Systems Design: the tracking and disturbance rejection problems", *IEEE Trans. on Auto. Cont.*, Vol. AC-26, pp. 203-217, (1981).
17. Sacks, R., and J. Murray, "Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem", unpublished notes, Texas Tech Univ., 1980.
18. Viyasagar, M., Schneider, H., and B. Francis, "Algebraic and Topological Aspects of Feedback System Stabilization", Tech Rpt. 80-90, Dept. of Elec. Engrg., Univ. of Waterloo, 1980.
19. Vidyasagar, M., and M. Viswanadham, "Algebraic Design Techniques for Reliable Stabilization", Tech Rpt. 81-02, Dept. of Elec. Engrg., Univ. of Waterloo, 1980.
20. Wolovitch, W.A., "Multivariable System Synthesis with Step Disturbance Rejection", *IEEE Trans. on Auto. Cont.*, Vol. AC-19, pp. 127-130, (1974).
21. Wolovitch, W.A., and P. Ferreira, "Output Regulation and Tracking in Linear Multivariable Systems", *IEEE Trans. on Auto. Cont.*, Vol. AC-24, pp. 460-465, (1979).
22. Youla, D.C., "Interpolary Multichannel Spectral Estimation", Unpublished Notes, Polytechnic Inst. of New York (1979).
23. Youla, D.C., Bongiorno, J.J., and H.A. Jabr, "Modern Wiener-Hopf Design of Optimal Controllers, Part I", *IEEE Trans. on Auto. Cont.*, Vol. AC-21, pp. 3-15, (1976).
24. Youla, D.C., Bongiorno, J.J., and H.A. Jabr, "Modern Wiener-Hopf Design of Optimal Controllers, Part II", *IEEE Trans. on Auto. Cont.*, Vol. AC-21, pp. 319-338, 1976.
25. Youla, D.C., Bongiorno, J.J., and C.N. Lu, "Single-Loop Feedback Stabilization of Linear Multivariable Dynamical Plants", *Automatica*, Vol. 10, pp. 159-173, (1974).
26. Zames, G., "Feedback and Optimal Sensitivity: Model reference transformation, weighted seminorms, and approximate inverses", *IEEE Trans. on Auto. Cont.*, (to appear).

A Design Method for Two-Dimensional Recursive Digital Filters

JOHN J. MURRAY, MEMBER, IEEE

Abstract—A method for designing two-dimensional, symmetric half-plane recursive digital filters is presented: a filter is first designed as a parametrized family of one-dimensional filters; a simple approximation is then used to find a rational, stable, two-dimensional filter. Some advantages and disadvantages of the method are discussed, and several examples are given.

I. INTRODUCTION

IN THE past few years, several successful algorithms for the design of recursive multidimensional digital filters have been published (e.g., [1]–[8]). However, it is probably safe to say that nothing approaching a “universal” design procedure has been developed. Of course, it is highly unlikely that such a procedure exists, in view of the fact that in the (much simpler) one-dimensional situation it is already clear that a multiplicity of design methods is necessary to handle the variety of problems which arise.

Although there are relatively few two-dimensional situations in which recursive filters are required, there are many in which recursive filters may be considered as an alternative to FIR filters. A major drawback in the application of recursive filters, however, has been the complexity and expense of their design, especially when one is merely exploring the possibility that they may be more efficient than FIR filters. Indeed, there is evidence that in many situations the FIR filters will be the more efficient [9]. For these reasons, most designers simply use FIR filters rather than go through the nonlinear optimization procedures usually necessary to design recursive filters and to ensure stability.

A gap thus appears in the range of design algorithms; there appear to be few procedures which can provide a relatively quick, suboptimal design which is guaranteed to be stable. This gap is highlighted by the fact that in two-dimensional filtering applications it is relatively rare to encounter a situation where it is important to meet precise specifications of the type encountered in one-dimensional filtering, e.g., limits on the passband ripple or the stopband attenuation. It is far more common to be given an idea of the “shape” of the filter without any numerical specifications. The design method described in this paper is directed towards this situation.

The basic element in this method is a recursive symmetric half-plane filter. This is in contrast with the asymmetric half-plane filters designed in [4] and the semirecursive symmetric

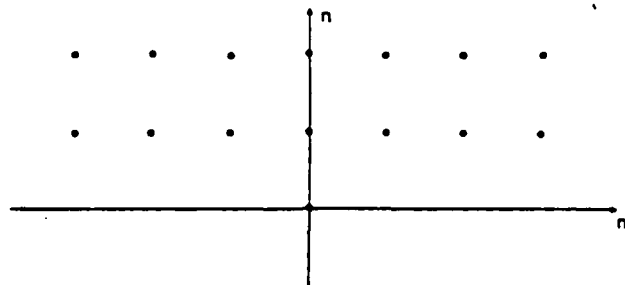


Fig. 1. Support set for symmetric half-plane filter.

half-plane filters in [10]. Preliminary versions of the present method have been described in [11] and [12]. It may be outlined as follows. Beginning with a frequency specification on the torus

$$T^2 = \{(\theta_1, \theta_2) \mid -\pi < \theta_1 \leq \pi, -\pi < \theta_2 \leq \pi\}$$

one first fixes θ_1 , and designs a classical one-dimensional recursive filter in the variable θ_2 . Doing this for each θ_1 gives a two-dimensional filter whose coefficients are (in general transcendental) functions of θ_1 . One then approximates these coefficient functions by trigonometric polynomials in θ_1 . The result is a rational symmetric half-plane two-dimensional filter. The major problem lies in carrying out the approximation in such a way that stability is preserved. We will describe an approximation procedure which is noniterative, requires a minimum of computation, and yields filters which can be proved to be stable.

II. SYMMETRIC HALF-PLANE FILTERS

By a recursive symmetric half-plane filter, we will mean a filter whose denominator is of the form

$$B(Z_1, Z_2) = 1 + \sum_{n=1}^N \sum_{m=-M}^M b_{mn} Z_1^m Z_2^n. \quad (1)$$

The corresponding support set is shown in Fig. 1. One advantage of these filters from the point of view of implementation is immediately clear; they are ideally suited for parallel processing since the output at any point depends only on outputs from the previous lines. It is therefore possible to process entire lines in parallel.

Another advantage of this class of filters is that its “stability set” is slightly smaller than that of asymmetric half-plane filters and, of course, considerably smaller than that of quarter-plane filters. The stability set for filters of the form (1) is the set defined by

$$S = \{(Z_1, Z_2) \in \mathbb{C}^2 \mid |Z_1| = 1 \text{ and } |Z_2| \leq 1\}.$$

Manuscript received June 9, 1980, revised July 13, 1981. This work was supported by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

The author is with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

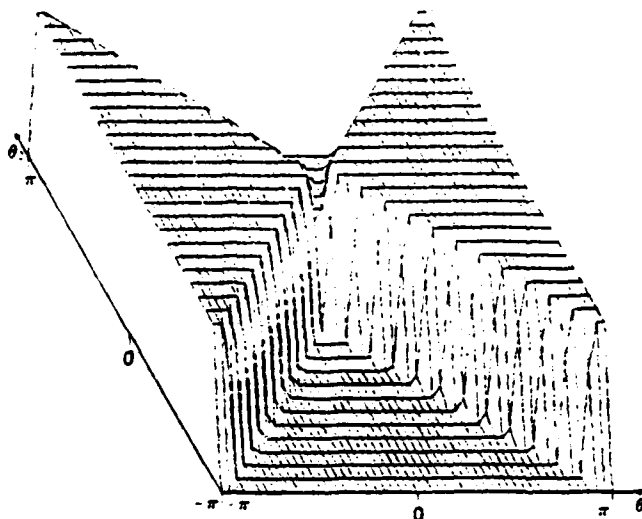


Fig. 2. A 90° fan filter.

(That is, if the denominator (1) does not vanish on the set S , the corresponding filter will be stable.) This is an immediate consequence of the corresponding fact for asymmetric half-plane filters [4] since the extra one-dimensional stability condition for the asymmetric filter is automatically satisfied in the present case—the function in question is identically 1. We also note that stability of the two-dimensional symmetric half-plane filter with transfer function equal to $H(Z_1, Z_2)$ is equivalent to stability of the entire family of one-dimensional filters (in Z_2) whose transfer functions are

$$H(e^{j\theta_1}, Z_2), \quad -\pi < \theta_1 < \pi.$$

This fact follows immediately from the form of the stability set S ; it will be important in the design procedure presented below.

Filters of the form (1), however, are incapable of approximating a general magnitude response. The restrictions which must be placed on a magnitude function $k(\theta_1, \theta_2)$ to enable it to be approximated by a stable symmetric half-plane filter (with constant numerator) may be stated as follows. The average gain over the set $\theta_1 = \text{constant}$, i.e.,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln k(\theta_1, \theta_2) d\theta_2$$

must be a constant independent of θ_1 . A simple direct proof of this condition is given in Appendix I.

It is clear from the form of these restrictions that in order to approximate a general magnitude response, one must cascade a symmetric half-plane filter with a one-dimensional filter in Z_1 . If one realizes this latter filter as a recursive filter, the cascade will be an asymmetric half-plane filter; if the one-dimensional filter is chosen to be a noncausal filter, the cascade will be essentially equivalent to a semirecursive symmetric half-plane filter. In both cases, the final filter is expressed as a product, rather than the more usual sum, of a one-dimensional filter and a symmetric half-plane filter. (This is not to imply that an arbitrary asymmetric half-plane or semirecursive symmetric half-plane filter can be factored in this way. However, the class of magnitude responses which can be approxi-

mated by the product form is the same as that associated with the sum form. In fact, it has been shown [13] that if one imposes the additional constraint that the magnitude response of a rational asymmetric half-plane filter has quadrantal symmetry then the filter can be expressed exactly as a product of the above type.)

One of the advantages of working with filters of the form (1) is the flexibility discussed above—by appropriate choice of the one-dimensional “compensating” filter one can realize either asymmetric filters or semirecursive filter. A further advantage is that any symmetries in the specification of the filter will be reflected in symmetries of the coefficients of the filter. Of course, other choices besides the two discussed above are possible for the one-dimensional filter; for instance, one could simply take it to be an FIR filter. This is in fact what is done in the examples presented below.

III. THE DESIGN PROCEDURE

For the sake of clarity we will describe the design procedure while simultaneously applying it to one particular filter specification—namely, the 90° fan filter shown in Fig. 2.

We assume, given a real frequency specification $h(\theta_1, \theta_2)$ and we wish to design a recursive filter, to approximate this specification. We proceed as follows.

1) For each value of θ_1 , we get a one-dimensional frequency specification

$$h_{\theta_1}(\theta_2) = h(\theta_1, \theta_2).$$

For the fan filter, this one-dimensional specification is an ideal low-pass filter with a cutoff frequency equal to $|\theta_1|$.

2) For each value of θ_1 , we use any of the design procedures available in one dimension to design a one-dimensional stable recursive filter to approximate $h_{\theta_1}(\theta_2)$ in the form

$$\prod_{k=1}^n \frac{a_k(\theta_1) + a_{k,1}(\theta_1)Z_2 + a_{k,2}(\theta_1)Z_2^2}{b_k(\theta_1) + b_{k,1}(\theta_1)Z_2 + b_{k,2}(\theta_1)Z_2^2}, \quad (2)$$

i.e., we design the one-dimensional filter as a cascade of first- and second-order sections.

For simplicity, we will take the bilinear transformation of a

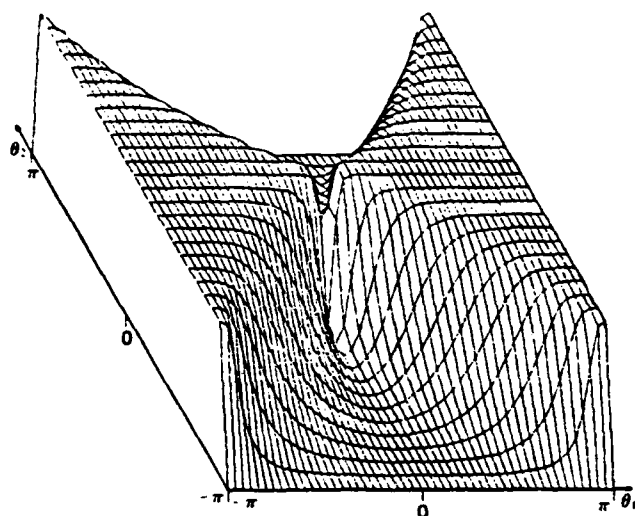


Fig. 3. Fan filter: Z_2 -order = 2, Z_1 -order = ∞ .

design the compensating filter from the ideal characteristic in (3); one must calculate the actual response of the two-dimensional filter, and design the compensating filter directly from this. Since it is not our purpose to go into one-dimensional filter design, we will not discuss this further.

Among the advantages of the above procedure is the fact that it can use classical one-dimensional filters, and so enables some of the intuition for and experience with these to be carried over to the two-dimensional case. For example, the ripple-free passbands and stopbands of the filters presented in the next section reflect the fact that the design is based on one-dimensional Butterworth filters.

A further advantage is that the response of the intermediate filter (which is rational in Z_2 and transcendental in Z_1) is easily calculated, and gives an upper bound on the performance of a filter with a given Z_2 -order. If this bound is not adequate, one must increase the Z_2 -order; this is significant in that increasing the Z_1 -order linearly increases the computation needed in implementation, while increasing the Z_2 -order linearly increases both computation and storage requirements. It is therefore normally preferable to increase the Z_1 -order rather than the Z_2 -order, if this is possible. In the present procedure, when one has decided on a Z_2 -order, increasing the Z_1 -order is simply a matter of computing more Fourier coefficients, and changing the window coefficients—one does not have to redo the bulk of the previous computation. Thus the procedure lends itself readily to interactive design.

IV. EXAMPLES

The first example is the fan filter designed in the previous section. Fig. 3 shows the response of the filter with Z_2 -order = 2 and Z_1 -order = ∞ . Fig. 4 shows the response of the final version. As calculated in the previous section, the denominator array is

$$\begin{array}{cccccccccc} 0.0000264 & 0 & 0.05252 & 0 & 0.2477 & 0 & 0.05252 & 0 & 0.0000264 \\ 0 & 0.001294 & 0 & -0.2653 & 0 & 0.2653 & 0 & 0.001294 & 0 \\ 1 \end{array}$$

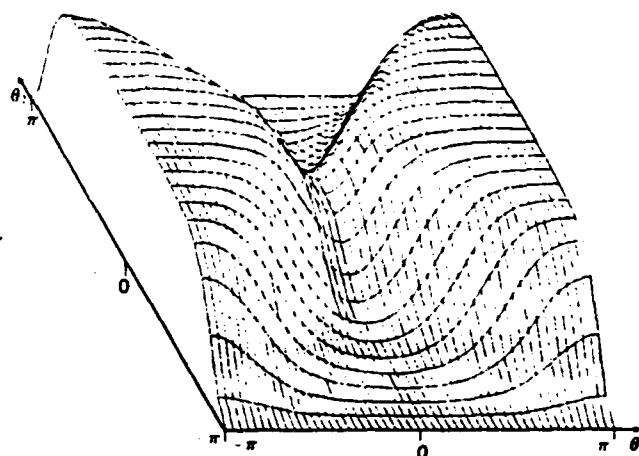


Fig. 4. Fan filter: single 2×9 section.

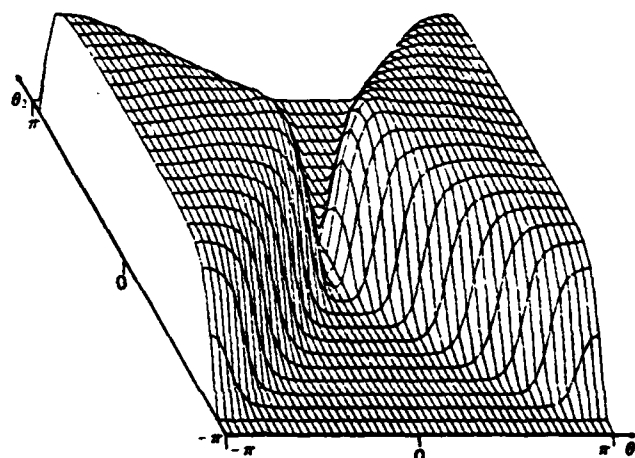


Fig. 5. Fan filter: cascade of four 2×17 sections.

while the numerator array is simply

$$\begin{array}{cccccccccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 1 \end{array}$$

The symmetry of the fan filter is clearly reflected in the symmetry of the denominator coefficients. While the resulting filter does not have an impressively sharp cutoff, it should be remembered that the two-dimensional part has only five distinct coefficients other than 0, 1, and 2, and was designed with a programmable calculator (SR-52). Total computation time was approximately 17 min with a 64-point quadrature for the Fourier coefficients. A higher-order (computer-calculated) version is shown in Fig. 5. The computation was not timed, but may be judged from the fact that the bulk of the computation consists of calculating forty one-dimensional Fourier coefficients. A circularly symmetric low-pass filter with cutoff frequency specified as $\pi/2$ is shown in Fig. 6; this is a cascade

second-order low-pass Butterworth filter in the present case. This gives

$$\frac{\omega_c^2(1+Z_1)^2}{\omega_c^2 + \sqrt{2}\omega_c + 1 + 2(\omega_c^2 - 1)Z_1 + (\omega_c^2 - \sqrt{2}\omega_c + 1)Z_1^2}$$

where

$$\omega_c = |\tan(\theta_1/2)|.$$

3) We express (2) in the form

$$\prod_{k=1}^n \frac{a_k(\theta_1)}{b_k(\theta_1)} \prod_{k=1}^n \frac{1 + \tilde{a}_{k,1}(\theta_1)Z_1 + \tilde{a}_{k,2}(\theta_1)Z_1^2}{1 + \tilde{b}_{k,1}(\theta_1)Z_1 + \tilde{b}_{k,2}(\theta_1)Z_1^2} \quad (3)$$

where

$$\tilde{a}_{k,i}(\theta_1) = a_{k,i}(\theta_1)/a_k(\theta_1)$$

and similarly for the denominator.

For the fan filter, we get

$$\frac{\omega_c^2}{1 + \sqrt{2}\omega_c + \omega_c^2} \cdot \frac{1 + 2Z_1 + Z_1^2}{1 + c(\theta_1)Z_1 + d(\theta_1)Z_1^2} \quad (4)$$

where

$$c(\theta_1) = \frac{2(\omega_c^2 - 1)}{1 + \sqrt{2}\omega_c + \omega_c^2}$$

and

$$d(\theta_1) = \frac{1 - \sqrt{2}\omega_c + \omega_c^2}{1 + \sqrt{2}\omega_c + \omega_c^2}.$$

At this stage, the first factor is just a one-variable function, and can, in principle, be approximated by a filter in Z_1 by using any suitable one-variable design algorithm. The second factor is a (transcendental) symmetric half-plane filter expressed as a cascade of second-order sections in Z_1 . Since each of the one-dimensional filters used is stable by design, it follows from the observation made in Section II that the resulting two-dimensional filter is stable, apart from possible singularities which may arise on the distinguished boundary. These can be shown to cause no stability problems in practice. (In the present example, such singularities occur at $\theta_1 = 0, \pi$; however, they disappear in the next step of the design procedure.)

4) The final step in the procedure is to approximate the functions $\tilde{a}_{k,i}(\theta_1)$ by trigonometric polynomials, and to approximate the functions $\tilde{b}_{k,i}(\theta_1)$ by trigonometric polynomials $\hat{b}_{k,i}(\theta_1)$ in such a way that stability is preserved. The latter approximation is the only one which presents any difficulties. It could be tackled by nonlinear optimization methods, but we wish to use a method which is less demanding from a computational point of view.

The most obvious approach is to window the Fourier series of the functions $\tilde{b}_{k,i}(\theta_1)$. It is shown in Appendix I that the resulting filter will be stable if the trigonometric kernel corresponding to the window used is positive everywhere and has total weight equal to one. However, for reasons discussed in Appendix II, this gives very large errors close to singularities, and the following slightly more complicated version is used instead.

Let

$$\alpha_k = \sqrt{1 + \tilde{b}_{k,2} + \tilde{b}_{k,1}}$$

and

$$\beta_k = \sqrt{1 + \tilde{b}_{k,2} - \tilde{b}_{k,1}}$$

and apply the windowing procedure to α_k and β_k (again with a positive kernel of weight one) to obtain trigonometric polynomials $\hat{\alpha}_k$ and $\hat{\beta}_k$. Finally, set

$$\hat{b}_{k,2} = \frac{1}{2}(\hat{\alpha}_k^2 + \hat{\beta}_k^2) - 1$$

and

$$\hat{b}_{k,1} = \frac{1}{2}(\hat{\alpha}_k^2 - \hat{\beta}_k^2).$$

It is shown in Appendix II that this procedure will yield a stable filter and can be expected to give much improved performance close to singularities. It should be noted, however, that this procedure approximately doubles the number of coefficients in the filter.

For the fan filter, we have

$$\alpha_1 = \left\{ \frac{4 \tan^2(\theta_1/2)}{\tan^2(\theta_1/2) + \sqrt{2} |\tan \theta_1/2| + 1} \right\}^{1/2}$$

and

$$\beta_1 = \left\{ \frac{4}{\tan^2(\theta_1/2) + \sqrt{2} \tan |\theta_1/2| + 1} \right\}^{1/2}.$$

Now since α_1 and β_1 are even, periodic functions of θ_1 , they can be expanded into Fourier cosine series. If we retain only the first three terms of each series (so that only three Fourier coefficients need actually be calculated for each), and apply a triangular window (Fejer kernel), we get

$$\hat{\alpha}_1 = 1.0587 - (0.2518) 2 \cos \theta_1 - (0.00514) 2 \cos 2 \theta_1$$

and

$$\hat{\beta}_1 = 1.0587 + (0.2518) 2 \cos \theta_1 - (0.00514) 2 \cos 2 \theta_1$$

and, so,

$$\hat{b}_{1,1} = 0.001294 Z_1^{-3} - 0.2653 Z_1^{-1} - 0.2653 Z_1 + 0.001294 Z_1^3$$

and

$$\hat{b}_{1,2} = 0.0000264 Z_1^{-4} + 0.05252 Z_1^{-2} + 0.2477 + 0.05252 Z_1^2 + 0.000264 Z_1^4.$$

This completes the design of the two-dimensional part of the filter since the numerator is already a function of Z_2 only.

As was mentioned previously, the one-dimensional compensating filter in the examples presented in the next section was chosen to be a simple FIR filter, which was designed by windowing. Actually, a separate compensating filter was designed for each second-order two-dimensional section [i.e., for each k , $1 \leq k \leq n$ in (2)], the order of the compensating filter being equal to the Z_1 -order of the corresponding section. Unfortunately, since the approximation procedure in step 4) exercises very poor control over the passband gain, it is not possible to

If we then let

$$\hat{s}(\theta_1) = \frac{1}{2\pi} \int_{-\pi}^{\pi} s(\phi) P(\theta_1 - \phi) d\phi$$

and

$$\hat{t}(\theta_1) = \frac{1}{2\pi} \int_{-\pi}^{\pi} t(\phi) P(\theta_1 - \phi) d\phi$$

\hat{s} and \hat{t} will be the windowed versions of s and t , and we will have

$$\begin{aligned} |\hat{s}(\theta_1)| &= \frac{1}{2\pi} \left| \int_{-\pi}^{\pi} s(\phi) P(\theta_1 - \phi) d\phi \right| \\ &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |s(\phi)| P(\theta_1 - \phi) d\phi \\ &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} (1 + t(\phi)) P(\theta_1 - \phi) d\phi \\ &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} 2P(\theta_1 - \phi) d\phi \\ &= 2. \end{aligned}$$

Since (5) is just $1 + \hat{t}(\theta_1)$, we get

$$|\hat{s}(\theta_1)| < 1 + \hat{t}(\theta_1) < 2$$

and so for each θ_1 , the polynomial

$$1 + \hat{s}(\theta_1)Z + \hat{t}(\theta_1)Z^2$$

has no zeros on the set $\{Z \mid |Z| \leq 1\}$. It follows immediately that the two-variable rational function

$$\frac{1}{1 + \hat{s}(\theta_1)Z + \hat{t}(\theta_1)Z^2}$$

is stable.

Q.E.D.

APPENDIX II

When the above windowing procedure was used directly, the resulting filters exhibited gross underdamping in regions where the bandwidth (in the θ_2 direction) was close either to 0 or 2π . We will now discuss the reason for this, and justify the procedure given in step 4) of the design.

If we are given a second order discrete time filter with denominator $1 + cZ + dZ^2$ (where c and d are arbitrary real numbers), we can transform it into a continuous-time filter by using the bilinear transform

$$Z = \frac{1-s}{1+s}.$$

This results in a second-order filter whose denominator is (with unity leading coefficient)

$$s^2 + 2 \frac{1-d}{1-c+d} s + \frac{1+c+d}{1-c+d}. \quad (6)$$

In general, the denominator of a second-order continuous-time transfer function can be expressed in the form

$$s^2 + 2\zeta\omega_0 s + \omega_0^2 \quad (7)$$

where ω_0 is the undamped natural frequency, and ζ the damping ratio of the system [17]. Comparison of (6) and (7) yields, for the continuous-time undamped natural frequency,

$$\omega_0 = \sqrt{\frac{1+c+d}{1-c+d}}$$

and for the damping ratio,

$$\zeta = \frac{1-d}{\sqrt{(1+c+d)(1-c+d)}}.$$

Thus, if the bandwidth approaches 0 or 2π , we must have

$$\text{either } 1+c+d \rightarrow 0$$

$$\text{or } 1-c+d \rightarrow 0$$

and if the damping ratio is to remain finite, we must have $1-d \rightarrow 0$ (this is the situation in the fan filter designed in Section III).

Now suppose the errors made in approximating c and d are of the same order, and that they swamp the actual values of the quantities which approach 0. If the order of error is ϵ , then

$$\zeta \sim \frac{\epsilon}{\sqrt{4\epsilon}} = \frac{\sqrt{\epsilon}}{2},$$

i.e., the damping ratio will be very small, and will decrease with increasingly accurate approximations to c and d . Thus not alone do we get poor performance, but increasing the order makes it worse! This is precisely what is observed in practice.

Suppose, however, that we approximate

$$\alpha = \sqrt{1+c+d} \text{ and } \beta = \sqrt{1-c+d}$$

and suppose for definiteness that $1-c+d \rightarrow 0$. Then in the above situation $d \rightarrow 1$ and $1+c+d \rightarrow 4$.

If $\hat{\alpha}$ and $\hat{\beta}$ denote the approximations, we then define

$$\hat{d} = \frac{1}{2} (\hat{\alpha}^2 + \hat{\beta}^2) - 1$$

and

$$\hat{c} = \frac{1}{2} (\hat{\alpha}^2 - \hat{\beta}^2)$$

and it then follows that the damping ratio of the approximated filter is

$$\hat{\zeta} = \frac{(2 - \hat{\alpha})(2 + \hat{\alpha}) + \hat{\beta}^2}{2\hat{\alpha}\hat{\beta}}.$$

If we make the same assumptions as before, we get $(2 - \hat{\alpha}) \sim \epsilon$ and $\hat{\beta} \sim \epsilon$ and, so,

$$\hat{\zeta} \sim \frac{\epsilon(4 - \epsilon) + \epsilon^2}{2\epsilon(2 - \epsilon)} \sim 1$$

—a much more acceptable result.

It remains only to show that this procedure gives a stable filter, i.e., that $|\hat{c}| < 1 + \hat{d} < 2$ for all θ_1 .

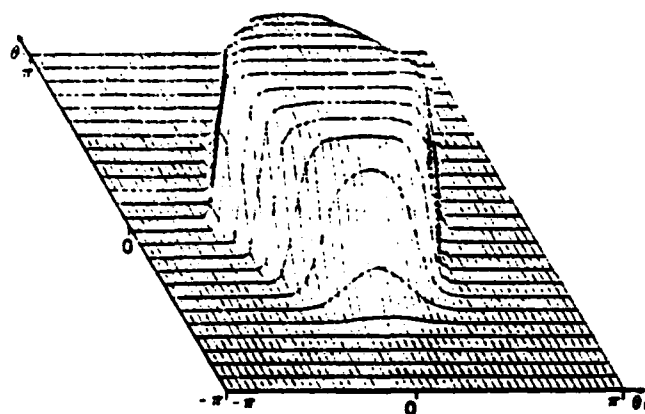


Fig. 6. Circularly symmetric low-pass filter: cascade of two 2×17 sections.

of two 2×17 filters. Finally, in order to indicate the numerical stability of the procedure, the same specification was used for the impractically large filter in Fig. 7. (Because of the high Z_1 -order, the effect of the one-dimensional filter was inserted artificially in this figure; also the elliptical shape of the passband is an artifact of the plotting system.)

We note that in the above filters, a cascade of two 2×17 sections, for example, would be a 4×34 filter if it were expressed in direct form, and so would require twice as many coefficients to be stored and manipulated. For this reason, these filters should be implemented in cascade form.

V. CONCLUSIONS

We have described an inexpensive, suboptimal method of designing stable half-plane recursive filters as a cascade of filters, each of which is second-order in the principal direction of recursion. In the examples presented, analytic expressions were available for the cascade form of the one-dimensional filters used; however, even when this is not the case, continuation methods [14], [15] are ideally suited to the numerical calculation of the appropriate coefficients.

The major open problem associated with the method is undoubtedly the development of more effective approximation methods for the denominator coefficient functions. While the method used above gives quite satisfactory results, one can not help feeling that one should be able to do as well with half the number of coefficients (since the approximating polynomial is squared after its order has been chosen). Furthermore, if the approximation procedure controlled the passband gain more tightly than the present method does, it would be possible to design the one-dimensional compensating filter from the ideal response rather than from the actual two-dimensional response. Thus an improved approximation procedure would yield multiple benefits.

Such approximation procedures, together with extensions of the above design method, are currently under investigation.

APPENDIX I

Here, we prove two statements made in the paper. The first is that if $k(\theta_1, \theta_2)$ is the magnitude response of any stable transfer function

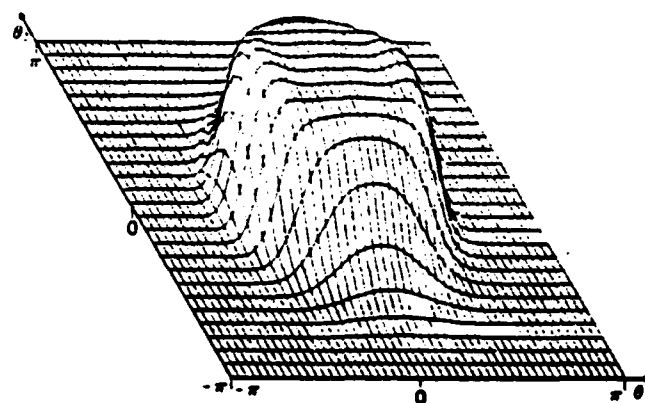


Fig. 7. Circularly symmetric low-pass filter: cascade of four 2×41 sections.

$$\frac{K}{B(Z_1, Z_2)} = \frac{K}{1 + \sum_{n=1}^N \sum_{m=-M}^M b_{mn} Z_1^m Z_2^n}$$

then

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln k(\theta_1, \theta_2) d\theta_2 = \ln K.$$

This follows from the fact that, by the stability condition, for each fixed θ_1 the one-variable polynomial

$$B(e^{j\theta_1}, Z)$$

has no zeros in the unit disk $\{Z | |Z| < 1\}$. It therefore has an analytic logarithm there. Therefore the function

$$\ln |B(e^{j\theta_1}, Z)| = \operatorname{Re} \ln B(e^{j\theta_1}, Z)$$

is harmonic on the unit disk and so, since the mean value of a harmonic function on a circle is equal to its value at the center of the circle, we get

$$\begin{aligned} \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln |B(e^{j\theta_1}, e^{j\theta})| d\theta &= \ln |B(e^{j\theta_1}, 0)| \\ &= \ln |1| \\ &= 0. \end{aligned}$$

Q.E.D.

The second statement to be proved is that windowing by a window whose kernel is positive and of total weight one preserves stability.

Proof: Let the kernel be $P(\theta)$ (an n th order trigonometric polynomial) and assume that $P(\theta) \geq 0, \forall \theta$, and

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} P(\theta) d\theta = 1.$$

Let $1 + s(\theta_1)Z + r(\theta_1)Z^2$ be any second-order polynomial in Z whose coefficients are periodic functions of θ_1 .

Suppose the function $1 + s(\theta_1)Z + r(\theta_1)Z^2$ has no zeros on the set $\{Z | |Z| \leq 1\}$ for each θ_1 . This is equivalent to [16]:

$$|s(\theta_1)| < 1 + r(\theta_1) < 2, \quad \forall \theta_1.$$

This is clearly equivalent to

$$|\hat{\alpha}^2 - \hat{\beta}^2| < \hat{\alpha}^2 + \hat{\beta}^2 < 4$$

and the first inequality is obvious from the fact that both $\hat{\alpha}^2$ and $\hat{\beta}^2$ are positive, while the second is a consequence of the following argument. We have

$$\begin{aligned}\hat{\alpha}^2 &= \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \alpha(\theta - \phi) \sqrt{P(\phi)} \sqrt{P(\phi)} d\phi \right\}^2 \\ &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} \alpha^2(\theta - \phi) P(\phi) d\phi \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\phi) d\phi \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \{1 + c(\phi) + d(\phi)\} P(\theta - \phi) d\phi\end{aligned}$$

with a similar inequality for $\hat{\beta}^2$. Adding gives

$$\hat{\alpha}^2 + \hat{\beta}^2 \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} 2\{1 + d(\phi)\} P(\theta - \phi) d\phi < 4. \quad \text{Q.E.D.}$$

REFERENCES

- [1] H. Chang and J. K. Aggarwal, "Design of two-dimensional recursive filters by interpolation," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 281-291, June 1977.
- [2] J. M. Costa and A. N. Venetsanopoulos, "Design of circularly symmetric two-dimensional recursive filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 432-443, Dec. 1974.
- [3] D. E. Dugeon, "Two-dimensional recursive filter design using differential correction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 264-267, June 1975.
- [4] M. P. Ekstrom and J. W. Woods, "Two dimensional spectral factorization with applications in recursive digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 115-128, Apr. 1976.
- [5] D. Goodman, "A design technique for circularly symmetric low-pass filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 290-304, Aug. 1978.
- [6] K. Hirano and J. K. Aggarwal, "Design of two-dimensional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 1066-1076, Dec. 1978.
- [7] G. A. Maria and M. M. Fahmy, "An L_p design technique for two-dimensional recursive filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 15-21, Feb. 1974.
- [8] S. Treitel and J. L. Shanks, "The design of multistage separable planar filters," *IEEE Trans. Geosci. Electron.*, vol. GE-9, pp. 10-27, Jan. 1971.
- [9] R. E. Twogood and M. P. Ekstrom, "Why filter recursively in two dimensions?," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1979, pp. 20-23.
- [10] D. B. Harris, "Design and implementation of rational two-dimensional digital filters," Ph.D. dissertation, Mass. Inst. Tech., Cambridge, MA, 1979.
- [11] J. Murray, "Symmetric half-plane filters," in *Proc. 20th Midwest Circuits Syst.*, Aug. 1977, vol. 1, pp. 434-436.
- [12] —, "A design method for 2-D digital filters," in *Proc. 13th Annu. Asilomar Conf. Circuits, Syst. Comput.*
- [13] D. M. Goodman, "Quadrantal symmetry conditions for nonsymmetric half-plane filters," Preprint UCRL-82443, Lawrence Livermore Lab., Mar. 1979.
- [14] K. S. Chao and R. Saeks, "Continuation methods in circuit analysis," *Proc. IEEE*, vol. 65, pp. 1187-1194, 1977.
- [15] E. Wasserstrom, "Numerical solutions by the continuation methods," *SIAM Rev.*, vol. 15, pp. 89-119, 1973.
- [16] E. I. Jury, *Inners and Stability of Dynamical Systems*. New York: Wiley, 1974.
- [17] M. B. Scherba, "Servo systems," in *Electronics Designers' Handbook*, 2nd ed., L. J. Giacoletto, Ed. New York: McGraw-Hill, 1977.



John J. Murray (M'78) was born in Galway, Ireland, on August 8, 1947. He received the B.Sc. and M.Sc. degree from University College, Cork, Ireland, in 1969 and 1970, respectively, and the Ph.D. degree from the University of Notre Dame, Notre Dame, IN, in 1974, all in mathematics.

He is currently with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX. His principal research interests are in the areas of several complex variables, multi-dimensional system theory, and time-varying systems.

245

Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem

RICHARD SAEKS, FELLOW, IEEE, AND JOHN MURRAY, MEMBER, IEEE

Abstract—An explicit relationship between the fractional representation approach feedback system design and the algebro-geometric approach to system theory is formulated and used to derive a global solution to the feedback system problem. These techniques are then applied to the simultaneous stabilization problem, yielding a natural geometric criterion for a set of plants to be simultaneously stabilized by a single compensator.

I. INTRODUCTION

CLASSICALLY, in control theory one is given a plant and desires to design a control system around this plant which meets certain design specifications. In fact, however, a "real world" plant is never known exactly and, as such, a realistic design must simultaneously meet specifications over an entire range of plants which (hopefully) include the actual plant. The simplest form of the resultant *simultaneous design problem* is the *robust design problem* wherein one desires to meet the design specifications in an ϵ ball around a prescribed nominal plant. Although this is satisfactory for dealing with modeling errors, it cannot cope with plants containing unknown parameters and/or plants characterized by multiple modes of operation. For instance, the dynamics of an airplane or rocket vary widely with altitude, while the dynamics of an electric motor change with speed and load. To cope with these problems, we must formulate a *simultaneous design theory* in which one designs a control system to simultaneously meet specifications over a prescribed set of plants. Of course, the set of plants may be taken to be a ball, in which case the classical robustness theory is replicated. Alternatively, one may choose to work with a set of plants in which one or more parameters vary over a prescribed range and/or a discrete set of plants, say, the dynamics of a two-speed motor in its high- and low-speed settings.

The simultaneous design concept is possibly best illustrated in the first-order case wherein a simple geometric solution suggests itself. Assume that our plants are of the form

$$p(s) = \frac{A}{s+B} \quad (1.1)$$

Manuscript received December 5, 1980; revised November 9, 1981. Paper recommended by E. W. Kamen, Chairman of the Linear Systems Committee. This work was supported in part by the Joint Services Electronics Program at Texas Tech University under ONR Contract 76-C-1136.

The authors are with the Department of Electrical Engineering, Texas Tech University, Lubbock, TX 79409.

and we desire to design a stable feedback system using a proportional compensator with gain t . This results in a system with characteristic function

$$d(s) = s + (B + tA) \quad (1.2)$$

and, as such, the feedback system will be stable if and only if $B + tA > 0$. Here, for a given compensator t , the feedback system will be stable if and only if the point (A, B) lies above the line with slope $1/t$ as shown in Fig. 1(a). As such, if we want to simultaneously stabilize an entire set of plants, their representations on the A - B plane must all lie above a line through the origin. For instance, the set of plants indicated by the hatched region in Fig. 1(b) can be simultaneously stabilized (by a compensator with gain $-1/2$), while the set of plants shown in Fig. 1(c) cannot be simultaneously stabilized since they subtend an angle greater than 180° on the A - B plane. Similarly, the set of plants shown in Fig. 1(d) cannot be simultaneously stabilized since they cross the negative A axis.

The above example suggest two alternative criteria for the simultaneous stabilization problem. One may adopt an algebraic criterion to the effect that

$$B + tA > 0 \quad (1.3)$$

for each plant in the prescribed set and some t . While such a test is definitive, it is local in nature, allowing one to test for stabilizability on a plant-by-plant basis, but yielding no global criterion with which to characterize a set of plants which is simultaneously stabilizable. To the contrary, one may adopt a global geometric viewpoint to the effect that a prescribed set of plants is simultaneously stabilizable if and only if it is contained in an appropriate half-plane. The goal of the present paper is the formulation of a similar geometric criterion for the simultaneous stabilization problem applicable to general linear systems.

The starting point for our theory is the ring-theoretic fractional representation theory introduced by the authors in a series of recent papers [9], [16] in which the set of compensators for a given plant are parameterized. Moreover, as a first cut at the simultaneous stabilization problem, one can reverse the role of the plant and compensator in this theory to parameterize the set of plants which are stabilized by a given compensator. In practice, however, one is not given a compensator *a priori* and, as such, we must characterize the set of plants obtained by the latter

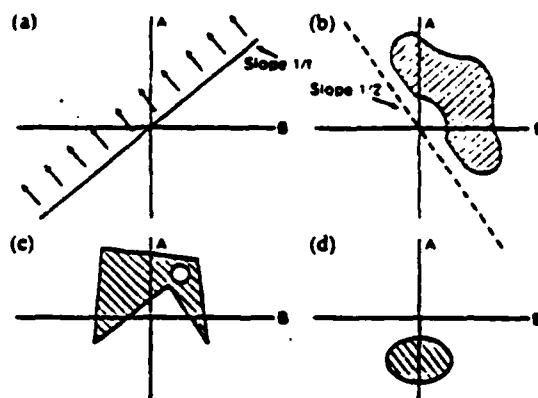


Fig. 1.

parameterization independently of the choice of compensator. For instance, in our first-order example, the set of plants in Fig. 1(a) are stabilized by a given compensator with slope $1/1$, while a given set of plants is simultaneously stabilizable if and only if it lies in the half-plane above some line through the origin. For the general problem, this is achieved by translating the fractional representation theory into an appropriate geometric setting in which the "shape" of the set of plants obtained from the latter parameterization may be characterized. In turn, the simultaneous stabilization problem may be resolved by requiring that the given set of plants lie in a region of the appropriate "shape."

Indeed, the appropriate geometric setting proves to be just the Grassmannian first introduced into the system theory literature by Hermann and Martin [11], [12]. Unlike their frequency domain formulation, however, we obtain the Grassmannian directly from the ring-theoretic fractional representation previously employed by the authors. Indeed, the Grassmannian is obtained simply by factoring out the nonuniqueness inherent in the fractional representation theory. As such, in addition to formulating the global theory necessary for our study of the simultaneous stabilization problem, the geometric approach yields new insight into the relationship between the fractional representation theory (which we identify with the elements of a general linear group) and the system itself (which we identify with the elements of a Grassmannian).

In the following section, the fractional representation theory is reviewed and the required Grassmannian is constructed. The resultant theory is then used to formulate a global description of the set of stabilizing compensators for a given plant in Section III. The resultant formulation also yields new insight into the problem of stabilizing a plant with a stable compensator [25] for which a necessary and sufficient condition is also derived in Section III. Finally, the simultaneous stabilization problem is investigated in Section IV wherein both global geometric and local algebraic criteria for the simultaneous stabilization of a prescribed family of plants are obtained.

Although the present paper is formulated in terms of the abstract (pseudo) coprime fractional representation theory of [5], [9], [10], [16]–[19], it should be pointed out that this

is simply one manifestation of a family of related approaches to the control system design problem developed during the past half-dozen years by Pernebo and Astrom [13], [14]; Antsaklis, Pearson, and Cheng [1], [6], [7]; Youla, Bongiorno, Jabr, and Lu [22]–[25]; and Zames [26] among others [2], [20], [21]. Indeed, the approaches of these authors are all closely related, any one of which could have been used as the basis for the present investigations. In particular, the formulation of Zames is applicable in a general ring-theoretic setting, and is essentially equivalent to that employed herein.

II. FRACTIONAL REPRESENTATION AND THE GRASSMANNIAN

The algebraic fractional representation theory is set in a nest of rings, groups, and multiplicative structures:

$$g \supset h \supset i \supset j.$$

Here, g is a ring with identity which represents the general class of systems with which we wish to work: rational matrices, continuous operators, a class of transcendental functions, etc.; and h is a subring of g containing the identity which models the systems which are stable in some sense: poles in a prescribed region, transcendental functions with restricted singularities, causal operators, etc. Finally, i denotes the multiplicative set composed of elements of h which admit an inverse in g , while j denotes the multiplicative subgroup of i made up of elements which are invertible in h . Detailed examples of this structure were given in [4] and [8] and will not be repeated here.

We say that a system $s \in g$ has a *right fractional representation* in $\{g, h, i, j\}$ if

$$s = n_{rr} d_{rr}^{-1} \quad (2.1)$$

where $n_{rr} \in h$ and $d_{rr} \in i$. Furthermore, we say that this representation is *right coprime* if there exist u_{rr} and v_{rr} in h such that

$$u_{rr} n_{rr} + v_{rr} d_{rr} = 1. \quad (2.2)$$

This equality is equivalent to the classical coprimeness concept for rational functions and matrices, while being defined in our general ring-theoretic setting. In particular, if g is the ring of rational functions and h is the ring of polynomials, (2.2) implies that n_{rr} and d_{rr} have no common zeros; and if g is the ring of rational functions and h is the ring of exponentially stable rational functions, (2.2) implies that n_{rr} and d_{rr} have no common right half-plane zeros.

Since g is, in general, noncommutative, we also define a *left fractional representation* for s via the equality

$$s = d_{ll}^{-1} n_{ll} \quad (2.3)$$

for $n_{ll} \in h$ and $d_{ll} \in i$. Furthermore, we say that this representation is *left coprime* if there exist u_{ll} and v_{ll} in h such that

$$n_{ll} u_{ll} + d_{ll} v_{ll} = 1. \quad (2.4)$$

Of course, in the classical case of a rational function or matrix, these fractional representations are assured to exist [10]. However, this is not the case in the general ring-theoretic setting. Therefore, for distributed, time-varying, and multidimensional systems, we assume that our plants admit such a representation as a prerequisite to the theory.¹ Interestingly, however, if such a representation exists, we may, without loss of generality, choose $u_{s/}$ and $v_{s/}$ such that the equality

$$v_{s/}u_{s/} = u_{s/}v_{s/} \quad (2.5)$$

also holds [8]. In this case, we say that the representation for s defined by the eight-tuple $\{n_{s/}, d_{s/}, u_{s/}, v_{s/}, n_{s/}, d_{s/}, u_{s/}, v_{s/}\}$ is *doubly coprime* and we express the defining equalities of (2.1)–(2.5) via the matrix equality $R_s^{-1} = S_s$, where

$$R_s = \begin{bmatrix} d_{s/} & -u_{s/} \\ n_{s/} & v_{s/} \end{bmatrix} \quad (2.6)$$

and

$$S_s = \begin{bmatrix} v_{s/} & u_{s/} \\ -n_{s/} & d_{s/} \end{bmatrix}. \quad (2.7)$$

It is interesting to compare the above formulation with that proposed by Zames [26]. Rather than working with an unstable system, s , Zames assumes that the given plant is first stabilized via classical techniques, and then develops his design theory around the resultant stable plant \tilde{s} . Now, since \tilde{s} is stable, it admits the *trivial right coprime representation* $\tilde{s} = \tilde{s}1^{-1}$ with the equality

$$[0][s] + [1][s] = u_{s/}n_{s/} + v_{s/}d_{s/} = 1 \quad (2.8)$$

implying right coprimeness, while a similar equality defines a left coprime representation for \tilde{s} . As such, the matrices $R_{\tilde{s}}$ and $S_{\tilde{s}}$ take on a very special form, permitting Zames to implement his design theory without explicitly dealing with u 's and v 's nor even introducing the coprimeness concept.

The key to our geometric formulation of the fractional representation theory lies with the observation that the 2×2 matrices R_s and S_s constitute a natural and concise representation for the given plant which can serve in lieu of the eight-tuple of n 's, d 's, u 's, and v 's. Indeed, if the input-output relation for our system is characterized by the equality

$$v_s = s\epsilon_s = [n_{s/}, d_{s/}^{-1}]\epsilon_s, \quad (2.9)$$

then the admissible input-output pairs [9] (ϵ_s, v_s) for our plant are parameterized by the equality

$$\begin{bmatrix} \epsilon_s \\ v_s \end{bmatrix} = \begin{bmatrix} d_{s/} & -v_{s/} \\ n_{s/} & v_{s/} \end{bmatrix} \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = R_s \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} \quad (2.10)$$

¹Computationally, the evaluation of $u_{s/}$ and $v_{s/}$ (or $u_{s/}$ and $v_{s/}$) reduces to the solution of a linear equation in the ring \mathcal{A} . In particular, if \mathcal{A} is a ring of stable rational functions or matrices thereof, one can multiply (2.2) and (2.4) by a common denominator, and thereby reduce their solution to that of a classical polynomial equation.

where $\sigma_s = [d_{s/}^{-1}]\epsilon_s$ is an appropriate "partial state" variable. As such, the 2×2 matrix R_s defines a natural model for the given system. Indeed, when such a model is employed, one can drop the invertibility requirement on $d_{s/}$ (and $d_{s/}$), although the matrix R_s must still admit an inverse with entries in \mathcal{A} .²

Since the 2×2 matrices R_s and S_s have entries in \mathcal{A} and admit an inverse which also has entries in \mathcal{A} , they form a group which we denote by $GL_{\mathcal{A}}(2)$, i.e., the general linear group of 2×2 matrices with entries in \mathcal{A} . If the elements of $GL_{\mathcal{A}}(2)$ are, however, to serve as a viable system representation, we must be cognizant of the fact that several such matrices may represent the same plant. The appropriate equivalence classes may, however, be characterized with the aid of the subgroup $E \subset GL_{\mathcal{A}}(2)$ composed of the upper triangular matrices

$$E = \begin{bmatrix} e_{11} & e_{12} \\ 0 & e_{22} \end{bmatrix} \quad (2.11)$$

in $GL_{\mathcal{A}}(2)$. Note that since these triangular matrices are assumed to be in $GL_{\mathcal{A}}(2)$, it follows that e_{11} and e_{22} are in \mathcal{A} .

Property 1: Let R_s and \tilde{R}_s be in $GL_{\mathcal{A}}(2)$. Then R_s and \tilde{R}_s represent the same system if and only if there exists $E \in E$ such that $R_s = \tilde{R}_s E$.

Proof: If $R_s = \tilde{R}_s E$ for some $E \in E$, then

$$\begin{bmatrix} \epsilon_s \\ v_s \end{bmatrix} = R_s \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = \tilde{R}_s E \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = \tilde{R}_s \begin{bmatrix} e_{11}\sigma_s \\ e_{22} \cdot 0 \end{bmatrix} = \tilde{R}_s \begin{bmatrix} \tilde{\sigma}_s \\ 0 \end{bmatrix} \quad (2.12)$$

where $\tilde{\sigma}_s = e_{11}\sigma_s$. As such, the set of input-output pairs defined by \tilde{R}_s coincides with those defined by R_s , except for a change of parameterization.

If R_s and \tilde{R}_s define the same set of input-output pairs, then for any such pair (ϵ_s, v_s) , there exist σ_s and $\tilde{\sigma}_s$ such that

$$\begin{bmatrix} \epsilon_s \\ v_s \end{bmatrix} = R_s \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = \tilde{R}_s \begin{bmatrix} \tilde{\sigma}_s \\ 0 \end{bmatrix} \quad (2.13)$$

which, in turn, implies that

$$\begin{bmatrix} \tilde{\sigma}_s \\ 0 \end{bmatrix} = \tilde{R}_s^{-1} R_s \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix} = E \begin{bmatrix} \sigma_s \\ 0 \end{bmatrix}. \quad (2.14)$$

Now, $E = \tilde{R}_s^{-1} R_s \in GL_{\mathcal{A}}(2)$ since R_s and \tilde{R}_s are in $GL_{\mathcal{A}}(2)$, while $R_s = \tilde{R}_s E$ by construction. As such, it suffices to show that $E \in E$. This, however, follows from the fact that (2.14) holds for all σ_s (and a corresponding $\tilde{\sigma}_s$). ■

Given that any two representations in $GL_{\mathcal{A}}(2)$ for the same system differ by a left factor in E , a natural setting for our system theory is the quotient space $GL_{\mathcal{A}}(2)/E$. Although $GL_{\mathcal{A}}(2)$ is a group and E is a subgroup, E is not normal and, as such, $GL_{\mathcal{A}}(2)/E$ is not a group. Fortunately, however, the resultant coset space (of equivalence classes) is a well-known and much studied geometric ob-

²In the case where $d_{s/}^{-1}$ does not exist, the plant defines a relation rather than a function on the input-output space. The resultant relation is, however, parameterized by σ_s . Moreover, since R_s is invertible, the relation is normal in the same sense of [15].

ject, the Grassmannian [1] $G_A(1,2)$, which we will adopt as the basic setting for our system theory.³

Since E is not a normal subgroup of $GL_A(2)$, $G_A(1,2)$ is not a group and, as such, does not admit an "internal" algebraic structure. The group $GL_A(2)$ does, however, act as a set of transformations on $G_A(1,2)$. Indeed, if $T \in GL_A(2)$ and $[U] \in G_A(1,2)$ is the equivalence class of $U \in GL_A(2)$, we define

$$T[U] = [TU] \in G_A(1,2). \quad (2.15)$$

Now, if $[U] = [V]$, then Property 1 implies that there exists $E \in E$ such that $V = UE$; hence,

$$T[V] = T[UE] = [TUE] = [TV] \quad (2.16)$$

via Property 1. As such, the operation of $GL_A(2)$ on $G_A(1,2)$ is well defined.

As a prerequisite to the formulation of our stabilization theory, it is necessary to characterize the stable systems as interpreted in $GL_A(2)$ and $G_A(1,2)$. Recall that if s is stable ($s \in k$), then s admits the doubly coprime fractional representation

$$R_s = \begin{bmatrix} 1 & 0 \\ s & 1 \end{bmatrix}. \quad (2.17)$$

Denoting the set of such 2×2 matrices by $W \subset GL_A(2)$, it then follows from Property 1 that the set of all representations for the stable systems in $GL_A(2)$ take the form $S = WE$ and, as such, they are represented by $[S] = [WE] = [W] \in G_A(1,2)$. Although W and E are both subgroups of $GL_A(2)$, they are not commutative and, as such, $S = WE$ is not a group. S is, however, characterized by the following property.

Property 2: Let

$$T = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}$$

be in $GL_A(2)$. Then $T \in S$ if and only if $t_{11} \in j$.

Proof: Since $S = WE$ if $T \in S$, then $T = WE$ for some $W \in W$ and $E \in E$; hence,

³If k is the field of scalars, $G_A(1,2)$ is the classical Grassmannian (of lines in 2-space), whereas if k is taken to be $n \times n$ matrices, $G_A(1,2)$ reduces to the classical Grassmannian of n planes in $2n$ space [1]. Although these classical Grassmannians have an analytic structure (compact manifold) which may not be shared by our abstract Grassmannian, the algebraic properties of $G_A(1,2)$ are all that is required for the present theory, and hence no difficulty is encountered by working with elements taken from an abstract ring. Indeed, for our purposes, we only use the fact that $G_A(1,2)$ is the coset space $GL_A(2)/E$ and identify it as the Grassmannian only to make the connection with the classical literature.

It is interesting to note that the Grassmannian has been used as a natural setting for multivariate systems by mathematical system theorists for a number of years [28], [29], [32]. Here, a system represented by an $n \times n$ frequency response matrix is identified with a curve taking values in the classical Grassmannian of n planes in $2n$ space. As such, that theory identifies a system with a Grassmannian-valued function, while our formulation identifies the system with a Grassmannian built from a ring of functions. Of course, the two approaches are completely equivalent in the multivariate case, while the present formulation is also well defined for general linear systems (time-varying, distributed, etc.).

$$\begin{aligned} T &= \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ w & 1 \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} \\ 0 & e_{22} \end{bmatrix} \\ &= \begin{bmatrix} e_{11} & e_{12} \\ we_{11} & (we_{12} + e_{22}) \end{bmatrix} \end{aligned} \quad (2.18)$$

showing that $t_{11} = e_{11} \in j$ [since E is invertible in $GL_A(2)$]. Conversely, if $t_{11} \in j$, we may factor T via

$$T = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ t_{21}t_{11}^{-1} & 1 \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} - t_{21}t_{11}^{-1}t_{12} \end{bmatrix} = WE. \quad (2.19)$$

Since $t_{11} \in j$, $w = t_{21}t_{11}^{-1}$ and $t_{22} - t_{21}t_{11}^{-1}t_{12}$ are in k . Thus, the unit triangular nature of W implies that it is in $GL_A(2)$. This, in turn, however, implies that

$$E = W^{-1}T \in GL_A(2) \quad (2.20)$$

since $GL_A(2)$ is a group. Since E is upper triangular and $E \in GL_A(2)$, $E \in E$, as was to be shown. ■

Since W and E are both groups, the equality $S = WE$ implies that $S^{-1} = EW$. Now, a little algebra similar to that used in the proof of Property 2 will reveal that $T \in S^{-1}$ if and only if $t_{22} \in j$. Indeed, S and S^{-1} are precisely the two classes of matrices for which the 2×2 matrix inversion formula is applicable [16].

Before concluding this section, it is instructive to comment on the relationship between $GL_A(2)$ and $G_A(1,2)$. In essence, $GL_A(2)$ is a set of representations for our systems, while $G_A(1,2)$ represents the set of systems. That is to say, $GL_A(2)$ is composed of the computationally tractable objects with which we actually describe a system, although many such objects may represent the same system. On the contrary, each element of $G_A(1,2)$ is uniquely identified with a system, and hence we may think of $G_A(1,2)$ as "being the set of systems" (or, at least, being in one-to-one correspondence with the set of systems). On the other hand, the elements of $G_A(1,2)$ are not computationally tractable, except through the intermediary of $GL_A(2)$. In practice, therefore, a plant is characterized by one of its representations in $GL_A(2)$, while the goal of the system design problem is to specify an appropriate compensator in $G_A(1,2)$. That is, we are designing a compensator rather than a representation of a compensator, and hence even though we work in $GL_A(2)$ as a matter of computational necessity, the result of the design process is an element of $G_A(1,2)$.

III. STABILIZATION

The basic feedback system we consider is shown in Fig. 2. The system is characterized by the connection equations

$$e_p = \mu_2 + v_c \quad (3.1)$$

and

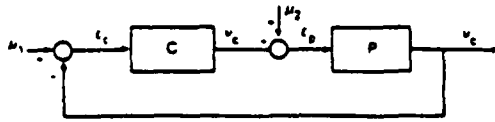


Fig. 2.

$$e_c = \mu_1 - v_p, \quad (3.2)$$

while the plant and compensator are characterized by

$$\begin{bmatrix} e_p \\ v_p \end{bmatrix} = R_p \begin{bmatrix} \sigma_p \\ 0 \end{bmatrix} \quad (3.3)$$

and

$$\begin{bmatrix} e_c \\ v_c \end{bmatrix} = R_c \begin{bmatrix} \sigma_c \\ 0 \end{bmatrix}, \quad (3.4)$$

respectively, where R_p and R_c are in $GL_k(2)$. Letting Q be the 2×2 matrix in $GL_k(2)$ defined by

$$Q = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad (3.5)$$

the connection equations (3.1) and (3.2) can be expressed as

$$\begin{bmatrix} e_p \\ v_p \end{bmatrix} = Q \begin{bmatrix} e_c \\ v_c \end{bmatrix} + \begin{bmatrix} \mu_2 \\ \mu_1 \end{bmatrix}. \quad (3.6)$$

A little algebra will also reveal that $Q^{-1} = Q' = -Q$; hence, this connection matrix is readily manipulated. Finally, the substitution of (3.3) and (3.4) into (3.6) yields the equality

$$\begin{aligned} \begin{bmatrix} \mu_2 \\ \mu_1 \end{bmatrix} &= R_p \begin{bmatrix} \sigma_p \\ 0 \end{bmatrix} - Q R_c \begin{bmatrix} \sigma_c \\ 0 \end{bmatrix} = R_p P_1 \begin{bmatrix} \sigma_p \\ \sigma_c \end{bmatrix} - Q R_c Q P_2 \begin{bmatrix} \sigma_p \\ \sigma_c \end{bmatrix} \\ &= [R_p P_1 + Q' R_c Q P_2] \begin{bmatrix} \sigma_p \\ \sigma_c \end{bmatrix}. \end{aligned} \quad (3.7)$$

Here $P_1 = \text{diag}[1, 0]$ and $P_2 = \text{diag}[0, 1]$.

On the basis of the above formulation, we say that the system is *stable* if and only if

$$[R_p P_1 + Q' R_c Q P_2] \in GL_k(2). \quad (3.8)$$

Since $[R_p P_1 + Q' R_c Q P_2]$ has entries in \mathcal{H} , (3.8) implies that its inverse exists and also has entries in \mathcal{H} . As such, a feedback system will be stable if and only if the relationship between its input vector $\text{col}(\mu_2, \mu_1)$ and its partial state vector $\text{col}(\sigma_p, \sigma_c)$ is stable. With the aid of (3.3) and (3.4), this, in turn, implies that the relationship between the system input vector and all of its internal variables is stable, and the converse is also true [9].

With these preliminaries, we now proceed with our first theorem in which a geometric characterization of the set of stabilizing compensators for a given plant is obtained. To this end, recall that S is the subset of $GL_k(2)$ corresponding to the stable systems and $[S]$ is its image in $G_k(1, 2)$, while any $T \in GL_k(2)$ defines a transformation on $[S] \subset G_k(1, 2)$ via

$$T[S] = \{[TS]; S \in S\} \subset G_k(1, 2). \quad (3.9)$$

Finally, let $R_c(R_p)$ denote the set of stabilizing compensators in $G_k(1, 2)$ for a given plant represented by $R_p \in GL_k(2)$.

Theorem: $R_c(R_p) = QR_p Q'[S]$.

Proof: To prove the theorem, we will show that the set of all representations for the stabilizing compensators in $GL_k(2)$ takes the form $QR_p Q'S$ for some $S \in S$, from which the theorem follows upon mapping these presentations into corresponding systems which are identified as elements of $G_k(1, 2)$. If $R_c = QR_p Q'S$ for some $S \in S$, then

$$\begin{aligned} [R_p P_1 + Q' R_c Q P_2] &= [R_p P_1 + Q'(QR_p Q'S)Q P_2] \\ &= R_p [P_1 + Q'S Q P_2]. \end{aligned} \quad (3.10)$$

Now, if

$$S = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \quad (3.11)$$

with $s_{11} \in j$, since $S \in S$, a little algebra will reveal that

$$Q'SQ = \begin{bmatrix} s_{22} & -s_{21} \\ -s_{12} & s_{11} \end{bmatrix} \in S^{-1} \quad (3.12)$$

since $(Q'SQ)_{22} = s_{11} \in j$. As such,

$$R_p [P_1 + Q'S Q P_2] = R_p \begin{bmatrix} 1 & -s_{21} \\ 0 & s_{22} \end{bmatrix} \in GL_k(2), \quad (3.13)$$

showing that the feedback system with compensator $QR_p Q'S$ is stable.

Conversely, if R_c is a stabilizing compensator for the feedback system,

$$[R_p P_1 + Q' R_c Q P_2] \in GL_k(2). \quad (3.14)$$

Moreover, since Q and R_p are in $GL_k(2)$, we may, without loss of generality, assume that R_c is of the form

$$R_c = QR_p Q'X, \quad (3.15)$$

in which case it suffices to show that

$$X = QR_p^{-1} Q' R_c \in S. \quad (3.16)$$

To this end, we observe that

$$\begin{aligned} [R_p P_1 + Q' R_c Q P_2] &= [R_p P_1 + Q'(QR_p Q'X)Q P_2] \\ &= R_p [P_1 + Q'X Q P_2] \\ &= R_p \begin{bmatrix} 1 & -x_{21} \\ 0 & x_{11} \end{bmatrix} \end{aligned} \quad (3.17)$$

since

$$X'XQ = \begin{bmatrix} x_{22} & -x_{21} \\ -x_{12} & x_{11} \end{bmatrix}. \quad (3.18)$$

Now, since $[R_p P_1 + Q' R_c Q P_2] \in GL_n(2)$,

$$\begin{bmatrix} 1 & -x_{21} \\ 0 & x_{11} \end{bmatrix} = R_p^{-1} [R_p P_1 + Q' R_c Q P_2] \in GL_n(2), \quad (3.19)$$

which implies that $x_{11} \in j$ as required to verify that $X \in S$. As such, an arbitrary stabilizing compensator for our system is of the form $R_c = QR_p Q' X$.

The set of all representations of all stabilizing compensators in $GL_n(2)$ thus takes the form $QR_p Q' S \subset GL_n(2)$. Now, upon mapping this set into the Grassmannian, we obtain

$$[QR_p Q' S] = QR_p Q' [S], \quad (3.20)$$

completing the proof. ■

Unlike the previous results given directly in terms of the fractional representation theory [9], [24] which are local in nature, the present theorem yields a global description of the set of stabilizing compensators for a given plant. Indeed, the required set is just a copy of the stable system in the Grassmannian transformed by the action of $QR_p Q'$.

It is interesting to compare the parameterization of the stabilizing compensators of the theorem with that obtained directly from the fractional representation theory. Indeed, a little algebra with the results of [4] will yield the equality

$$R_c = QR_p Q' W; \quad W \in W \quad (3.21)$$

for the family of stabilizing compensators. Recalling, however, that $S = WE$, this parameterization differs from that of the theorem only in that the equivalence transformation E has been deleted. As such, the set of stabilizing compensators of (3.21) includes exactly one representation for each stabilizing compensator rather than parameterizing all representations for the stabilizing compensators as is the case with the present theorem. Of course,

$$[QR_p Q' W] = QR_p Q' [W] = QR_p Q' [S], \quad (3.22)$$

showing that the same set of compensators in the Grassmannian is defined by the two theories.

It follows immediately from the theorem that a stabilizing compensator always exists given that the plant is modeled by an $R_p \in GL_n(2)$. In practice, however, one often requires that the compensator be of a special form: stable, memoryless, diagonal, etc. As such, if $[C]$ represents the desired class of compensators in $G_n(1,2)$, it is necessary to design a compensator in $[C] \cap QR_p Q' [S]$, in which case it is not clear that such a compensator even exists. In the case where $[C]$ represents the stable systems, i.e., we desire to find a stable stabilizing compensator [25], a simple criterion for the existence of the required compensator can, however, be obtained as follows.

Corollary 1: A plant represented by $R_p \in GL_n(2)$ can be stabilized by a stable compensator if and only if

$$R_p \in S^{-1} S.$$

Proof: If $R_p \in S^{-1} S$, there exist S_1 and S_2 in S such that

$$R_p = S_2^{-1} S_1. \quad (3.23)$$

Now let

$$R_c = QS_2^{-1} Q' \quad (3.24)$$

which is stable since the congruence transformation defined by Q maps S^{-1} to S and conversely. Moreover, since $Q' = Q^{-1}$,

$$\begin{aligned} R_c &= QS_2^{-1} Q' = QS_2^{-1} Q' Q S_1 Q' Q S_1^{-1} Q' \\ &= (QS_2^{-1} S_1 Q') (Q S_1^{-1} Q') = QR_p Q' S \end{aligned} \quad (3.25)$$

where $S = QS_1^{-1} Q' \in S$. As such, we have constructed a stable stabilizing compensator for R_p as required.

Conversely, if $R_c \in S$ is a stable stabilizing compensator for R_p , then by the theorem, $R_c = QR_p Q' S$ for some $S \in S$. As such,

$$\begin{aligned} R_p &= Q' R_c S^{-1} Q = Q' R_c Q Q' S^{-1} Q \\ &= (Q' R_c Q) (Q' S^{-1} Q) = S_2^{-1} S_1 \end{aligned} \quad (3.26)$$

since the congruence transformation defined by Q maps $R_c \in S$ to $S_2^{-1} \in S^{-1}$ and $S^{-1} \in S^{-1}$ to $S_1 \in S$. ■

Of course, when the hypotheses of the corollary hold, the set of stable stabilizing compensators for $R_p = S_2^{-1} S_1$ is nonvoid and given by $[S] \cap QR_p Q' [S]$. No explicit parameterization for this set is, however, known nor is it obvious that one even exists. This intuition will be explored further in the following section on the simultaneous stabilization problem.

IV. SIMULTANEOUS STABILIZATION

The key to our solution of the simultaneous stabilization problem lies with a reversal of the analysis used in the derivation of the stabilization theorem. That is, one assumes that $R_c \in GL_n(2)$ is given and parameterizes the set of plants which are stabilized by R_c . Denoting that set of plants by $R_p(R_c)$, we obtain the following theorem. Since the proof for the theorem is virtually identical to that already given for the stabilization theorem of the previous section, it will not be repeated here.

Theorem: $R_p(R_c) = QR_c Q' [S]$.

Since every $T \in GL_n(2)$ is of the form $T = QR_c Q'$ (with $R_c = Q' T Q$), the lemma implies that a set of plants $P \subset G_n(1,2)$ can be simultaneously stabilized if and only if they lie in a copy of the stable systems $[S]$ transformed by the action of some $T \in GL_n(2)$. Indeed, this is the desired generalization of the first-order example given in the introduction to the case of a general linear system. In the general case, the Grassmannian plays the role of the A - B plane, the stables, $[S]$, play the role of the half-plane, and the general linear group, $GL_n(2)$, serves as the group of "rotations." These observations are summarized in the following corollary.

Corollary 2: A set of plants $P \subset G_n(1,2)$ can be simultaneously stabilized if and only if they lie in $T[S]$ for some $T \in GL_n(1,2)$.

Although the corollary represents a complete geometric solution to the simultaneous stabilization problem, it is not amenable to convenient implementation. We therefore give an alternative algebraic solution to the problem in $GL_n(2)$.

Corollary 3: Let $R_p \in GL_n(2)$, $p \in P$ be representations for a set of plants $P \subset G_n(1,2)$. Then P can be simultaneously stabilized if and only if there exists a family of matrices $S_p \in S$, $p \in P$ such that

$$S_p^{-1}S_q = R_p^{-1}R_q; \quad p, q \in P.$$

Proof: If R_c simultaneously stabilizes R_p , $p \in P$, then it follows from Corollary 2 that there exist $S_p \in S$, $p \in P$ such that

$$R_p = QR_cQ'S_p; \quad p \in P. \quad (4.1)$$

Hence,

$$R_p^{-1}R_q = (QR_cQ'S_p)^{-1}(QR_cQ'S_q) = S_p^{-1}S_q. \quad (4.2)$$

Conversely if there exists $S_p \in S$, $p \in P$ such that $R_p^{-1}R_q = S_p^{-1}S_q$; $p, q \in P$, then we let p' be an arbitrary plant in P and define R_c via

$$R_c = Q'R_pS_p^{-1}Q. \quad (4.3)$$

R_c is independent of the choice of p' . Indeed, if p'' is an alternative choice, then since $R_{p'}^{-1}R_{p''} = S_{p'}^{-1}S_{p''}$,

$$R_{p'}S_{p'}^{-1} = R_{p'}(R_{p'}^{-1}R_{p''}S_{p''}^{-1}) = R_{p''}S_{p''}^{-1}. \quad (4.4)$$

Moreover, for any $p \in P$, $R_p^{-1}R_{p'} = S_p^{-1}S_{p'}$; hence,

$$\begin{aligned} R_p &= R_{p'}S_{p'}^{-1}S_p = QQ'R_{p'}S_{p'}^{-1}QQ'S_p \\ &= Q(Q'R_{p'}S_{p'}^{-1}Q)Q'S_p = QR_cQ'S_p \end{aligned} \quad (4.5)$$

from which it follows, via the corollary, that the set of plants $P \subset G_n(1,2)$ is simultaneously stabilized by R_c . ■

It is interesting to note that in the special case in which P contains exactly two plants, say p and q , then they are simultaneously stabilizable if and only if

$$R_p^{-1}R_q \in S^{-1}S \quad (4.6)$$

which is identical to the criterion of Corollary 1 to stabilize a plant with a stable compensator. Indeed, this should not be surprising since the problem of stabilizing a plant with a stable compensator is completely equivalent to the problem of simultaneously stabilizing the given plant and the zero plant [represented by $R_z = 1 \in GL_n(2)$]. This follows immediately from the definition of stability used in [4] or, alternatively, one may let $R_z = 1$ and

$$R_c = \begin{bmatrix} d_{cr} & -u_{cl} \\ n_{cr} & v_{cl} \end{bmatrix} \quad (4.7)$$

in which case

$$[R_1P_1 + Q'R_cQ'P_2] = \begin{bmatrix} 1 & -n_{cr} \\ 0 & d_{cr} \end{bmatrix}. \quad (4.8)$$

As such, R_c stabilizes the zero plant if and only if $d_{cr} \in j$ or, equivalently, $R_c \in S$. The problem of stabilizing p with a stable compensator is thus equivalent to simultaneously stabilizing p and the zero plant. Since the zero plant is represented by $1 \in GL_n(2)$, it follows from (4.6) that p can be stabilized by a stable compensator if and only if

$$R_p = (1)^{-1}R_p = R_z^{-1}R_p \in S^{-1}S \quad (4.9)$$

which replicates the result of Corollary 3.

Of course, a similar argument can be used to obtain an algebraic criterion for the simultaneous stabilization of a set of plants P by a stable compensator. To this end, we simply augment the set of plants P by the zero plant and then apply the theorem to $P' = \{0\} \cup P$. Alternatively, a geometric criterion for the simultaneous stabilization of P may be obtained by requiring that $P \subset T[S]$ for some $T \in S^{-1}$. Since the corresponding R_c is given by

$$R_c = Q'TQ, \quad (4.10)$$

the resultant compensator will be stable.

Finally, the possibility of parameterizing the set of compensators (or stable compensators) which simultaneously stabilize P should be considered. In effect, this amounts to parameterizing the set of 2×2 matrices $T \in GL_n(2)$ or $T \in S^{-1}$ such that $P \subset T[S]$ which, in turn, requires some kind of parameterization for P . In particular, if $P = T[S]$, then $R_c = Q'TQ$ is the unique stabilizing compensator, while the stabilizing compensators for a single plant p may be parameterized by $[S]$ as per the stabilization theorem of Section III.

V. EXAMPLES

Since the action of $GL_n(2)$ on $G_n(1,2)$ is a geometric invariant, one can, at least intuitively, say that the "shape" of $T[S]$ is identical to that of $[S]$, and every set whose "shape" is the same as $[S]$ may be obtained from $[S]$ by such a transformation. As such, Corollary 4 implies that a prescribed set of plants $P \subset G_n(1,2)$ admits a simultaneous stabilization if and only if P is contained in a subset of the Grassmannian whose "shape" is the same as $[S]$. For instance, in the example of Fig. 1, P must be contained in an appropriate half-plane.

Although well-defined coordinate systems do exist in the Grassmannian, at the time of this writing we have yet to formulate a computational algorithm for implementing the above-described simultaneous stabilization problem in any degree of generality. Interestingly, however, in the case where P is composed of exactly two plants (a two-speed motor or the dynamics of a swing-wing aircraft), a simple frequency domain criterion for the simultaneous stabilization problem was given by Chua [8], [16] and the authors in the single-variate case, and has since been extended to the multivariate case by Vidyasagar and Viswanadham [19].

To illustrate the above-described general geometric criterion in a computationally trackable setting, consider the

case of a set of third-order single-variate plants

$$p(s) = \frac{q_2 s^2 + q_1 s + q_0}{s^3 + p_2 s^2 + p_1 s + p_0} \quad (5.1)$$

which are to be stabilized by a proportional compensator $c(s) = t$. To represent the class of plants geometrically, we identify the six-dimensional space of such plants with a family of three-dimensional Euclidian spaces (with coordinates p_0, p_1, p_2) parameterized by the numerator coefficients q_0, q_1 , and q_2 . As such, we have a three-dimensional family of three-dimensional spaces $R^3_{(q_0, q_1, q_2)}$. Of course, if there is no feedback ($t = 0$), such a system will be stable if and only if

$$p_0, p_2 > 0 \quad (5.2)$$

and

$$p_1 > p_0/p_2 \quad (5.3)$$

independently of the numerator parameters. The resultant stability region for the $t = 0$ case is thus as illustrated in Fig. 3(a).

In the case of a nonzero compensator, a similar argument with the Hurwitz criterion will yield the set of inequalities

$$p_0 + tq_0 > 0 \quad (5.4)$$

$$p_2 + tq_2 > 0 \quad (5.5)$$

and

$$p_1 > (p_0 + tq_0)/(p_2 + tq_2). \quad (5.6)$$

As such, for a fixed value of the numerator coefficients q_0, q_1 , and q_2 , the new stability region is identical in "shape" to the $t = 0$ case except for a shift of the origin to the point (tq_0, tq_1, tq_2) in $R^3_{(q_0, q_1, q_2)}$, as illustrated in Fig. 3(b). Of course, as t changes, the origin of the stability region moves along the line determined by the point (q_0, q_1, q_2) . Unlike the $t = 0$ case, however, where the stability region is independent of the numerator parameters, the line along which the origin of the stability region moves is determined by the numerators parameters. Thus, in this example, our three-dimensional family of three-dimensional Euclidian spaces $R^3_{(q_0, q_1, q_2)}$ plays the role of the Grassmannian, with the region defined by (5.2) and (5.3) in each such space characterizing the stable systems. Furthermore, the real group (corresponding to the proportional compensators) acts on this space by translating the stability region in the space $R^3_{(q_0, q_1, q_2)}$ along a line determined by the point (q_0, q_1, q_2) .

Although the above example was derived from basic principles, we believe that it illustrates the essential geometric nature of the simultaneous stabilization problem as formulated in our abstract theory. Indeed, a prescribed set of plants can be simultaneously stabilized if and only if they are contained in a translate of the stables.

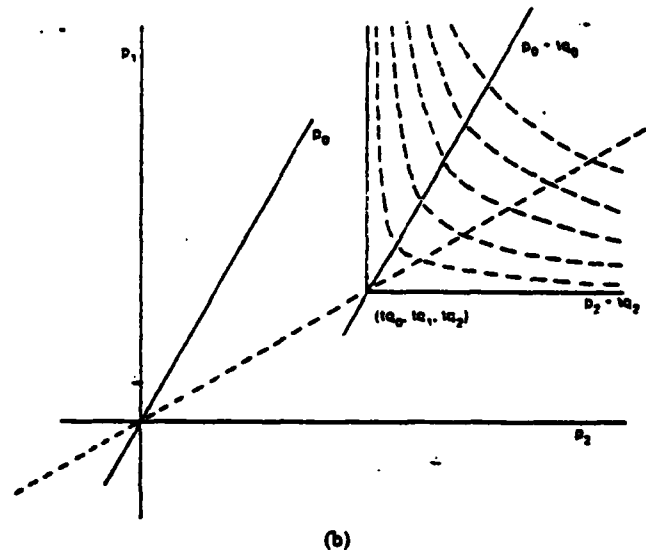
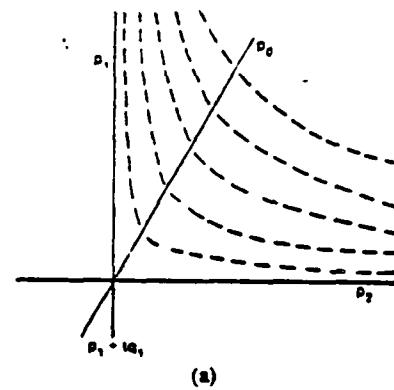


Fig. 3.

VI. CONCLUSIONS

Our purpose in the preceding has been threefold. First, we have attempted to exhibit the essential relationship between the fractional representation theory and the algebraic-geometric approach to system theory. Second, we have presented a global solution to the feedback system stabilization problem. Third, a solution to the simultaneous stabilization problem has been presented. It should, however, be pointed out that the solution presented for the simultaneous stabilization problem is mathematical in nature and not intended for computational implementation. At the present time, no computationally feasible solution to the simultaneous stabilization problem is known, except in the case where P contains exactly two plants wherein a simple frequency domain test is possible [8], [16] and in the simple low dimensional illustrated herein.

REFERENCES

- [1] P. J. Antsaklis and J. B. Pearson, "Stabilization and regulation in linear multivariable systems," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 928-930, 1978.
- [2] G. Bengtsson, "Output regulation and internal modes—A

- frequency domain approach," *Automatica*, vol. 13, pp. 333-345, 1977.
- [3] F. Brickell and R. S. Clark, *Differentiable Manifolds*. London, England: Van Nostrand Reinhold, 1970.
- [4] R. W. Brockett and C. I. Byrnes, "Multivariable Nyquist criteria, root locus, and pole placement: A geometric viewpoint," *IEEE Trans. Automat. Contr.*, to be published.
- [5] F. M. Callier and C. A. Desoer, "Stabilization, tracking and disturbance rejection in linear multivariable distributed systems," in *Proc. 17th IEEE Conf. Decision Contr.*, San Diego, CA, Jan. 1979, pp. 513-514; also, Univ. California, Berkeley, Tech. Memo. UCB/ERL M78/83, Dec. 1978.
- [6] L. Cheng and J. B. Pearson, "Frequency domain synthesis of multivariable linear regulators," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 3-15, 1978.
- [7] —, "Synthesis of linear multivariable regulators," *IEEE Trans. Automat. Contr.*, to be published.
- [8] O. Chua, M.S. thesis, Texas Tech Univ., Lubbock, 1980.
- [9] C. A. Desoer, R.-W. Liu, J. Murray, and R. Saeks, "Feedback system design: The fractional representation approach to analysis and synthesis," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 399-412, 1980.
- [10] B. Francis and M. Vidyasagar, "Algebraic and topological aspects of the servo problem for lumped linear systems," unpublished notes, Yale Univ., New Haven, CT, 1980.
- [11] R. Hermann and C. Martin, "Applications of algebraic geometry to linear system theory," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 19-25, 1977.
- [12] —, "Applications of algebraic geometry to system theory: The MacMillan degree and Kronecker indices as topological and holomorphic invariants," *SIAM J. Contr.*, vol. 16, pp. 743-755, 1978.
- [13] L. Pernebo, Ph.D. dissertation, Lund Inst. Technol., Lund, Sweden, 1978.
- [14] —, "An algebraic theory for design of controllers for linear multivariable systems: Parts I and II," *IEEE Trans. Automat. Contr.*, to be published.
- [15] R. Saeks, *Generalized Networks*. New York: Holt, Rinehart and Winston, 1972.
- [16] R. Saeks, J. J. Murray, O. Chua, and C. Karmokolias, "Feedback system design: The single variate case," unpublished notes, Texas Tech Univ., Lubbock, 1980.
- [17] R. Saeks and J. Murray, "Feedback systems design: The tracking and disturbance rejection problems," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 203-217, 1981.
- [18] M. Vidyasagar, H. Schneider, and B. Francis, "Algebraic and topological aspects of feedback system stabilization," *Dep. Elec. Eng., Univ. Waterloo, Waterloo, Ont., Canada, Tech. Rep. 80-09*, 1980.
- [19] M. Vidyasagar and M. Viswanadham, "Algebraic design techniques for reliable stabilization," *Dep. Elec. Eng., Univ. Waterloo, Waterloo, Ont., Canada, Tech. Rep. 81-02*, 1980.
- [20] W. A. Wolovich, "Multivariable system synthesis with step disturbance rejection," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 127-130, 1974.
- [21] W. A. Wolovich and P. Ferreira, "Output regulation and tracking in linear multivariable systems," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 460-465, 1979.
- [22] D. C. Youla, "Interpolary multichannel spectral estimation," unpublished notes, Polytechnic Inst. New York, Brooklyn, 1979.
- [23] D. C. Youla, J. J. Bongiorno, and H. A. Jabr, "Modern Wiener-Hopf design of optimal controllers, Part I," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 3-15, 1976.
- [24] —, "Modern Wiener-Hopf design of optimal controllers, Part II," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 319-338, 1976.
- [25] D. C. Youla, J. J. Bongiorno, and C. N. Lu, "Single-loop feedback stabilization of linear multivariable dynamic plants," *Automatica*, vol. 10, pp. 159-173, 1974.
- [26] G. Zames, "Feedback and optimal sensitivity: Model reference transformation, weighted seminorms, and approximate inverses," *IEEE Trans. Automat. Contr.*, to be published.



Richard Saeks (S'59-M'65-SM'74-F'77) was born in Chicago, IL, in 1941. He received the B.S. degree in 1964, the M.S. degree in 1965, and the Ph.D. degree in 1967 from Northwestern University, Evanston, IL, Colorado State University, Fort Collins, and Cornell University, Ithaca, NY, respectively, all in electrical engineering.

He is presently Paul Whitfield Horn Professor of Electrical Engineering, Mathematics, and Computer Science, Texas Tech University, Lubbock, where he is involved in teaching and re-

search in the areas of fault analysis, large-scale systems, and mathematical system theory.

Dr. Saeks is a member of AMS, SIAM, ASEE, and Sigma Xi.



John Murray (M'78) was born in Galway, Ireland, on August 8, 1947. He received the B.Sc. and M.Sc. degrees from University College, Cork, Ireland, in 1969 and 1970, respectively, and the Ph.D. degree from the University of Notre Dame, Notre Dame, IN, in 1974, all in mathematics.

He is currently with the Department of Electrical Engineering, Texas Tech University, Lubbock. His principal research interests are in the areas of several complex variables, multidimensional system theory, and time-varying systems.

SIMULTANEOUS DESIGN OF CONTROL SYSTEMS

by

R. Saeks and J. Murray
Department of Electrical Engineering
Texas Tech University
Lubbock, Texas 79409

ABSTRACT

The problem of designing a feedback controller which stabilizes a number of plants simultaneously is discussed from the fractional representation point of view. An abstract solution of this general simultaneous stabilization problem is presented, and an elementary, explicit criterion is given for the simultaneous stabilizability of two systems. Finally, some examples and counter examples are presented, and some open problems are discussed.

1. INTRODUCTION

Classically, in control theory one is given a plant and desires to design a control system around this plant which meets certain design specifications. In fact, however, a "real world" plant is never known exactly and, as such, a realistic design must simultaneously meet specifications over an entire range of plants which (hopefully) include the actual plant. The simplest form of the resultant *simultaneous design problem* is the *robust design problem* wherein one desires to meet the design specifications in an ϵ -ball around a prescribed nominal plant. Although this is satisfactory for dealing with modeling errors it cannot cope with plants containing unknown parameters and/or plants characterized by multiple modes of operation. For instance, the dynamics of an airplane or rocket vary widely with altitude while the dynamics of an electric motor change with speed and load. To cope with these problems we must formulate a *simultaneous design theory* in which one designs a control system to simultaneously meet specifications over a prescribed set of plants. Of course, the set of plants may be taken to be a ball in which case the classical robustness theory is replicated. Alternatively, one may choose to work with a set of plants in which one or more parameters vary over a prescribed range and/or a discrete set of plants; say, the dynamics of a two speed motor in its high and low speed settings.

The purpose of this paper is to review the state of research on the simultaneous design problem including the derivation of an explicit criterion of the simultaneous stabilization of two distinct plants and an algebro-geometric solution of the general simultaneous stabilization problem. In addition, the fundamental relationship between the simultaneous stabilization problem and the problem of designing a stable (or minimally

unstable) stabilizing compensator for a given plant is reviewed.

2. SIMULTANEOUS STABILIZATION AND STABLE STABILIZATION

We consider the feedback system shown in Fig. 1; this is characterized by the connection equations

$$e_p = u_2 + v_c$$

$$e_c = u_1 - v_p$$

or

$$\begin{bmatrix} e_p \\ v_p \end{bmatrix} = Q \begin{bmatrix} e_c \\ v_c \end{bmatrix} + \begin{bmatrix} u_2 \\ u_1 \end{bmatrix}$$

where

$$Q = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad (2.1)$$

To describe the dynamics of the plant and controller, we use the abstract fractional representation theory of [1], [2], [3]. This assumes four sets

$$g \supset h \supset i \supset j$$

where g is a ring with identity which represents the general class of systems with which we wish to work, h is a subring of g corresponding to the stable systems in g , i is a multiplicative set consisting of the elements in h which have inverses in g , and j is the subgroup of h which consists of the elements of h which have inverses in h .

We assume that the plant P has a right-coprime fractional representation

$$P = N_r D_r^{-1}$$

where the coprimeness is exhibited by

$$U_r N_r + V_r D_r = 1.$$

and also a left-coprime representation

$$P = D_L N_L^{-1}$$

with

$$N_L U_L + D_L V_L = 1.$$

It has been shown [2] that one can then find U_L and V_L so that, in addition,

$$V_L U_L = U_L V_L.$$

Thus the plant P can be described by the matrix

$$R_P = \begin{bmatrix} D_L & -U_L \\ N_L & V_L \end{bmatrix}$$

or its inverse

$$S_P = \begin{bmatrix} V_L & U_L \\ -N_L & D_L \end{bmatrix}.$$

The admissible input-output pairs (c_p, v_p) for the plant are then described by

$$\begin{bmatrix} c_p \\ v_p \end{bmatrix} = R_P \begin{bmatrix} \sigma_p \\ 0 \end{bmatrix}$$

where c_p is a kind of "partial state" for the plant P .

The corresponding matrices for the controller will be denoted by R_C and S_C , etc..

It can then be shown [3], [4] that a given controller C will stabilize a plant P if and only if

$$R_P = Q R_C Q^T W \quad (2.2)$$

where Q is the connection matrix (2.1), and W is of the form

$$\begin{bmatrix} 1 & 0 \\ S & 1 \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} \\ 0 & e_{22} \end{bmatrix} \quad (2.3)$$

where each matrix is in $GL_2(h)$ - i.e., is a 2×2 matrix of elements of h which has an inverse whose elements are in h . It is easy to see that being of the form (2.3) is equivalent to being an element of $GL_2(h)$ whose $(1,1)$ -element is in j , and that is in turn is equivalent to being the R -matrix of a stable system. Thus in terms of the R -matrix representation, we can restate (2.2) as

Theorem 1: [4]

The set of plants stabilizable by a compensator C is precisely the set of stable plants transformed by left multiplication by $Q R_C Q^T$.

This can also be restated as follows: A set P of plants is simultaneously stabilizable iff P lies in an image of the stables under left multiplication by some element of $GL_h(2)$.

These criteria, while geometrically appealing, can not be handled analytically; the following equivalent criterion is therefore useful.

Theorem 2: [4]

Let P be a set of plants represented by R_p , $p \in P$. The set P is simultaneously stabilizable iff there exists a family of matrices W_p , $p \in P$, of the form (2.3) such that

$$W_p^{-1} W_q = R_p^{-1} R_q, \quad \forall p, q \in P.$$

Proof:

The necessity of this condition is obvious from (2.2). The sufficiency can be checked by defining

$$R_C = Q^T R_p W_p^{-1} Q$$

for any $p \in P$, and using the condition to check that R_C is well defined. Then (2.2) follows.

One can insert the requirement that the compensator be stable merely by adjoining the zero plant to the given set of plants - a compensator is stable if and only if it stabilizes the zero plant. Thus the stable stabilization of n plants simultaneously can be treated as a problem of simultaneously stabilizing $n+1$ plants. The converse, although less obvious, is also true - see [3].

3. THE TWO-PLANT CASE

The only case of which we know in which even the analytic criterion in section 2 can be implemented is the case of two plants. In this case, the criterion is as follows: two plants, with representing matrices R_1 and R_2 , can be simultaneously stabilized iff there exist matrices W_1 and W_2 of the form (2.3) such that

$$W_1^{-1} W_2 = R_1^{-1} R_2 \quad (3.1)$$

We note that, since the identity R -matrix corresponds to the zero plant, this is equivalent to the condition that the "plant" $R_1^{-1} R_2$ have a stable stabilizing compensator. (This is an example of the converse mentioned at the end of section 2.) For the linear time-invariant case, the problem has been solved by Youla [5]. It is of interest, however, to relate Youla's solution

to our approach. To this end, we denote $R_1^{-1}R_2$ by

$$R = \begin{bmatrix} D & -U \\ N & V \end{bmatrix}$$

and restrict ourselves to the scalar-input scalar-output case. If we write (3.1) in the form

$$W_1 R = W_2$$

and use the fact that the (1,1)-entry in each of the W -matrices must be in j , we get the condition:

$$D + TN \in j \text{ for some } t \in h.$$

In the present case, this is clearly equivalent to requiring the existence of a stable, minimum-phase transfer function which, at each closed right half-plane zero of N , interpolates D to an order equal at least to the order of the zero of N . Continuity and realness of the transfer functions show that a necessary condition for this is that D have the same sign at all closed positive real-axis zeros of N . Youla showed that this was also sufficient.

Thus to solve the problem of simultaneously stabilizing two plants, it remains only to express N and D in terms of the original plants. An easy calculation shows that we can take

$$N = N_2 D_1 - N_1 D_2$$

and

$$D = V_1 D_2 + U_1 N_2$$

where

$$R_1 = \begin{bmatrix} D_1 & -U_1 \\ N_1 & V_1 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} D_2 & -U_2 \\ N_2 & V_2 \end{bmatrix}$$

Thus in order for the plants to be simultaneously stabilizable, it is necessary and sufficient that $V_1 D_2 + U_1 N_2$ have the same sign at all closed positive real-axis zeros of $N_2 D_1 - N_1 D_2$.

Some calculations involving the coprimeness conditions show that this is equivalent to the condition that D_2/D_1 have the same sign at all closed positive real-axis zeros of $N_2 D_1 - N_1 D_2$.

This gives a criterion in terms of the transfer-functions themselves [3][4].

4. EXAMPLES AND PROBLEMS

The only case in which the geometric results can be illustrated on paper is the case in which there are only two parameters. For this reason our first example will deal with this situation.

EXAMPLES: [4]

Suppose given a family of plants of the form

$$p(s) = \frac{A}{s+B}$$

we would like to know when it is possible to stabilize these simultaneously by use of a proportional compensator with gain t . In this case the denominator of the closed-loop transfer function is

$$d(s) = s + (B+tA)$$

and so the feedback system will be stable iff $B+tA > 0$. Thus a set of plants is simultaneously stabilizable iff for some t , $B+tA > 0$ for each plant in the set. Now each plant can be represented as a point in the (A,B) -plane; in this representation, a set of plants is simultaneously stabilizable iff the set lies entirely above some straight line through the origin. The slope of such a line is $-t$, where t is the gain of a stabilizing compensator. For example, the set in Figure 2.b is stabilizable, while the sets in Figures 2.c and 2.d are not. Since the set of stables in this case is the upper half-plane, this gives a very vivid (although very special) illustration of theorem 1: a set of plants is simultaneously stabilizable iff it is contained in a rotated (or more accurately, sheared) version of the set of stables.

Our second example is a counterexample to the effect that even if every pair of plants in a set is simultaneously stabilizable, the entire set may not be. To this end, we take a set consisting of three plants $\{p_0, p_1, p_2\}$; here p_0 is the zero plant, and

$$p_1 = n_1/d_1, \quad p_2 = n_2/d_2$$

where n_1, d_1, n_2, d_2 are graphed in Figure 3. It is easy to see, by using Youla's criterion, that p_1 and p_2 each have stable stabilizing compensators, and by using the criterion in section 3, that p_1 and p_2 have a simultaneous stabilizing compensator. However, there is no stable compensator which simultaneously stabilizes p_1 and p_2 . If there were, then by the criterion of Theorem 2 there would be a stable transfer function f such that both $d_1 + fn_1 = r_1$ and $d_2 + fn_2 = r_2$ were stable and minimum-phase. At the zeros of $n_1, d_1 > 0$ and so r_1 must be positive on the positive real axis. Similarly r_2 must be positive on the positive real axis. However, if we eliminate f we see that

$$n_2 d_1 - n_1 d_2 = n_2 r_1 - n_1 r_2$$

and so, at the zeros of $n_2 d_1 - n_1 d_2$, we must have $n_2 r_1 = n_1 r_2$. But at these points, $n_1 > 0$ and $n_2 < 0$, and so we get a contradiction. Thus p_0, p_1

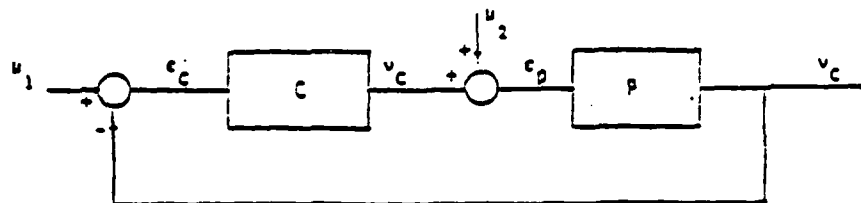


Figure 1. Feedback System

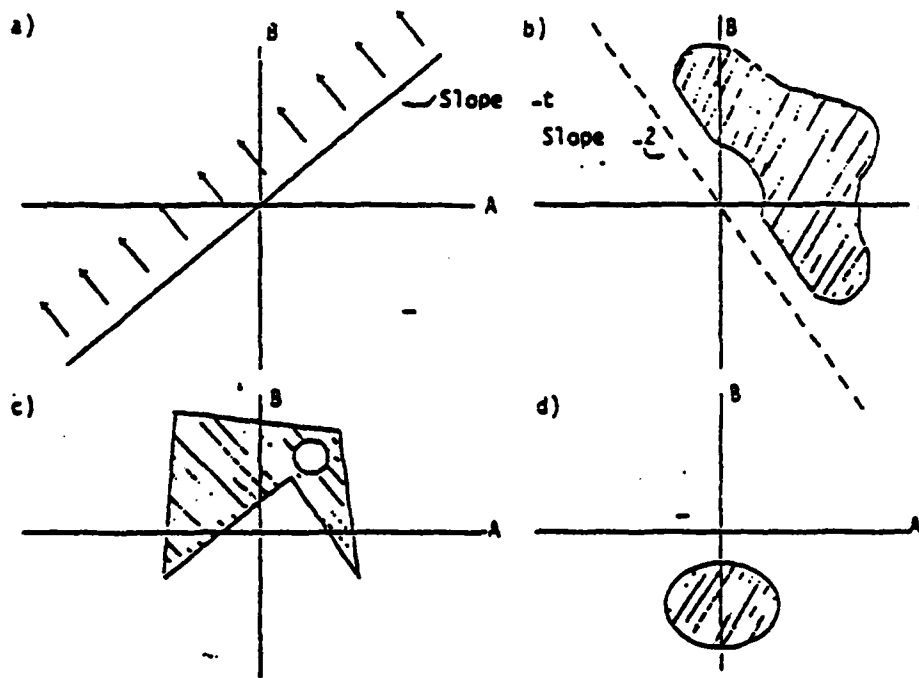


Figure 2. Simultaneous stabilization of a 1st order plant with a proportional compensator.

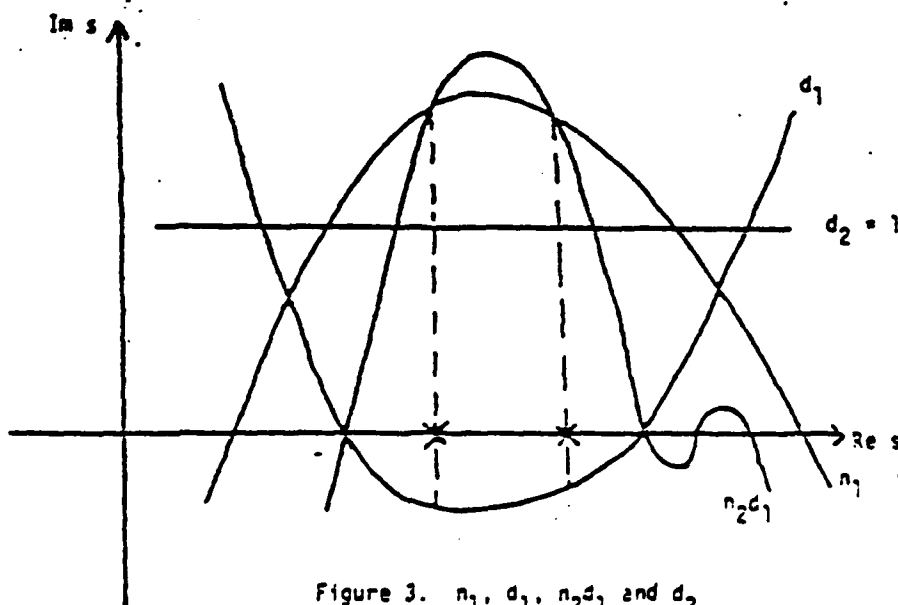


Figure 3. n_1 , d_1 , $n_2 d_1$ and d_2

and p_2 can not be simultaneously stabilizable.

It should be clear from the above that there are more problems than solutions in this area of research. At present, for example, we do not know any testable necessary and sufficient conditions for the simultaneous stabilizability of even three scalar plants. Thus as a first problem, we might state:

Problem 1: Find necessary and sufficient conditions for the simultaneous stabilizability of $N(>2)$ plants.

The following problem may also be of interest:

Problem 2: If one knows that a set of plants can be simultaneously stabilized, how does one find all compensators (or even some compensators) which will stabilize the set?

Problem 3: Since stabilization alone is usually not enough, can one find conditions for the existence of a compensator which, in addition to simultaneously stabilizing a set of plants, will also cause them to satisfy some other conditions (e.g. track a specified input signal)?

5. CONCLUSIONS

We have discussed the problem of finding a compensator which stabilizes every plant in a given set of plants. In the abstract, both geometric and analytic criteria have been given for the existence of such a compensator. However, only in very special cases, such as the case of two plants, can these criteria be checked. Another special case, in which the geometric criterion becomes particularly clear, is the case of first-order plants with proportional controllers. We have also given an example where pairwise simultaneous stabilizability of a set of plants does not imply overall simultaneous stabilizability. Finally, we have indicated some directions for further research.

REFERENCES

1. Desoer, C.A., Liu, R.-w., Murray, J., and R. Sacks, "Feedback System Design: the Fractional Representation Approach to Analysis and Synthesis," *IEEE Trans. Autom. Contr.*, **AC-25** (1980) 399-412.
2. Sacks, R., and J. Murray, "Feedback System Design: the Tracking and Disturbance Rejection Problems," *IEEE Trans. Autom. Contr.*, **AC-26** (1981) 203-217..
3. Vidyasagar, M., and M. Viswanadham, "Algebraic Design Techniques for Reliable Stabilization," Report 81-02, Dept. of Elec. Eng., University of Waterloo.
4. Sacks, R., and Murray, J., "Fractional Representation, Algebraic Geometry, and the Simultaneous Stabilization Problem," unpublished notes, Texas Tech Univ., 1980.
5. Youla, D.C., Bongiorno, J.J., and C.N. Lu, "Single-Loop Feedback Stabilization of Linear Multivariable Dynamical Plants," *Automatica*, **10** (1974), 159-173.

DETECTION AND ESTIMATION IN IMAGERY

26th

Signal recovery from signal dependent noise

Rangaschar Kasturi, Thomas F. Krile, John F. Walkup
Department of Electrical Engineering, Texas Tech University
Lubbock, Texas 79409

Abstract

It is well known that the noise processes corrupting an image are in general signal-dependent. An interesting aspect of signal-dependent noise is that there is a certain amount of signal-information embedded in the noise. Most of the image restoration techniques, however, attempt to suppress the noise terms to obtain an estimate of the image and do not exploit the additional signal information contained in the noise. A simple technique designed to demonstrate the potential for signal extraction from signal-dependent noise is presented in this paper.

Introduction

An interesting aspect of signal-dependent noise sources is that a certain amount of signal information is embedded in the noise.¹ Although the signal dependence of the noise renders the image restoration problem more complicated, better estimates should be possible if the additional information contained in the signal-dependent noise is extracted in the estimation process. In fact, in situations where the noise term dominates the signal term, the signal information present in the noise term may be larger than the signal term itself. A simple technique designed to recover the signal from signal-dependent noise is presented in this paper.

Signal recovery from Gaussian signal-dependent noise

Consider the image formation model² given by the following equation:

$$R = cS + kf(S)N_1 + N_2 \quad (1)$$

Here R is the noisy measurement, S is the signal to be estimated, $f(S)$ is a function of the signal, N_1 and N_2 are signal-independent, statistically independent zero mean Gaussian random processes with variances c_1^2 and c_2^2 respectively, and c and k are scalar constants. In order to simulate the "worst case" situation, we let $c=0$ and $k=1$ in Eq. (1) to obtain

$$R = f(S)N_1 + N_2 \quad (2)$$

In this model, all the signal information is contained in the signal-dependent noise. In this situation, estimation of the signal is possible only by exploiting the signal-dependence of the noise. The mean, μ_R , and the variance, c_R^2 , of the observation R are given by

$$\mu_R = E[f(S)N_1 + N_2] = 0 \quad (3)$$

and

$$\begin{aligned} c_R^2 &= E[(R - \mu_R)^2] \\ &= E[(f(S))^2]c_1^2 + c_2^2 \end{aligned} \quad (4)$$

At each observation point we may treat the signal S as an unknown constant. Then the variance of R conditioned on $S = s$ is

$$\sigma_{R|S=s}^2 = (f(s))^2 c_1^2 + c_2^2 \quad (5)$$

Then we have

$$s = f^{-1} \left\{ \sqrt{\frac{\sigma_{R|S=s}^2 - c_2^2}{c_1^2}} \right\} \quad (6)$$

Thus when the function describing the signal dependence has an inverse, an estimate of the signal, \hat{s} , may be obtained by estimating the variance of the noisy observation, $\hat{\sigma}_{R|S=s}^2$, and a knowledge of the parameters of the noise processes. In particular when

$$f(s) = s^p, \quad (7)$$

where p is a known non-zero constant, we obtain

$$s = \left\{ \frac{\hat{\sigma}_{R|S=s}^2 - \sigma_2^2}{\sigma_1^2} \right\}^{\frac{1}{2p}}. \quad (8)$$

An estimate of the variance of the noisy observation R may be obtained by assuming the random variable to be ergodic on a local basis and using a spatial window of size $M \times M$ centered around the pixel being estimated; i.e.,

$$\hat{\sigma}_{R_{i,j}}^2 = \frac{1}{M^2} \sum_k \sum_l \left[r_{(i+k), (j+l)} - \hat{\mu}_{R_{i,j}} \right]^2, \quad (9)$$

where

$$\hat{\mu}_{R_{i,j}} = \frac{1}{M^2} \sum_k \sum_l r_{(i+k), (j+l)}. \quad (10)$$

Here the subscripts i, j represent the coordinates of the pixel being estimated, $r_{i,j}$ is the noisy observation at i, j and the summation is carried out over all the pixels in the estimation window. In order to reduce the effect of spatial smoothing it is desirable to keep the size of the estimation window small. A noisy image obtained when the observation model of Eq. (2) is applied to the original image of Fig. 1, with $f(s) = \sqrt{s}$ is shown in Fig. 2. A constant background has been added to this image so that the resulting gray levels are all positive. (This constant background, however, will not affect the estimate obtained as the estimate is based on the variance.) The image extracted from this noisy observation using the estimation technique described in this section is shown in Fig. 3. This restored image indicates that the estimator is successful in recovering the gross structure of the original image.

The estimation technique presented above is not limited to the "worst case" imaging situation discussed here, and has been extended to situations when the signal embedded in the noise is only a part of the total signal information.³

Signal recovery from Poisson noise

Images formed at low light levels are corrupted by noise associated with the discrete nature of the photon counting process. Such images are well modeled⁴ by the equation,

$$R = \frac{1}{\lambda} P[\lambda S], \quad (11)$$

where $P[\lambda S]$ represents a Poisson process with λS as the parameter, and R and S are the noisy observation and the original signal respectively. The scaling factor $\frac{1}{\lambda}$ is used

to make the mean of the measurement R and the mean of the signal S equal to each other. Note that in this model the noise is inherently signal-dependent. A noisy image obtained from the original image of Fig. 1 using this model is shown in Fig. 4. When an estimate of the signal is subtracted from the noisy observation, the residual noisy image still contains signal information. As an example, consider the difference image, R_D , obtained

by subtracting the original signal S from the noisy observation R of Eq. (11). This difference signal is given by

$$R_D = \frac{1}{\lambda} P[\lambda S] - S. \quad (12)$$

The mean and variance of R_D are given by

$$\mu_{R_D} = E\left[\frac{1}{\lambda} P[\lambda S] - S\right] = 0, \quad (13)$$

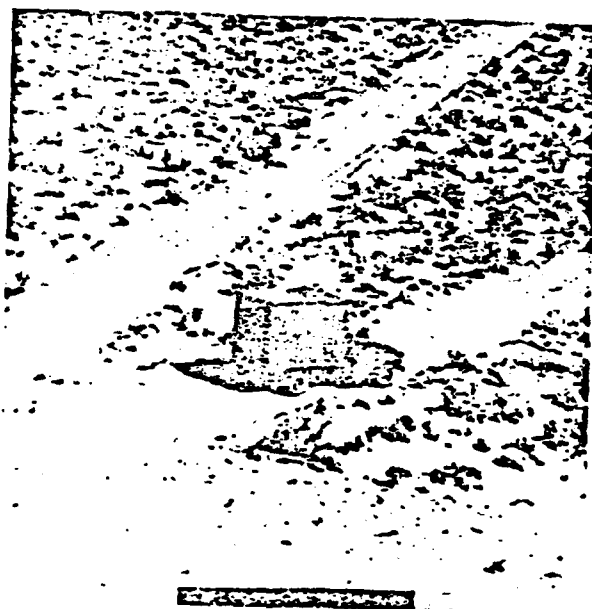


Figure 1. The original image.

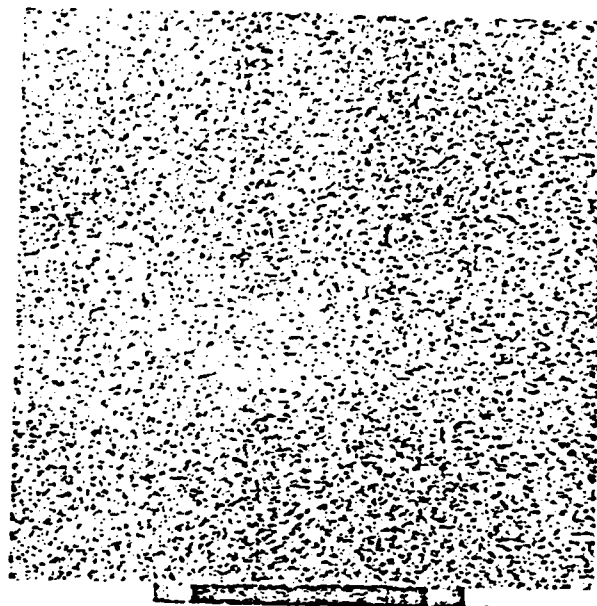


Figure 2. The noisy image with Gaussian noise, with $\sigma_1=4.0$ and $\sigma_2=20.0$.

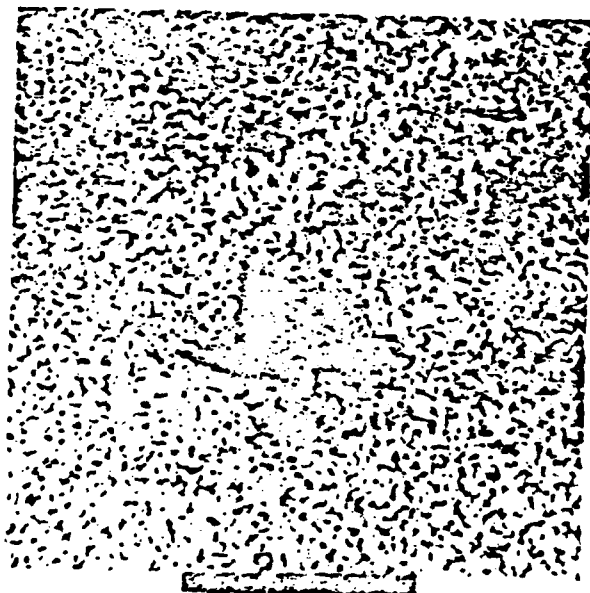


Figure 3. The image recovered from Fig. 2.

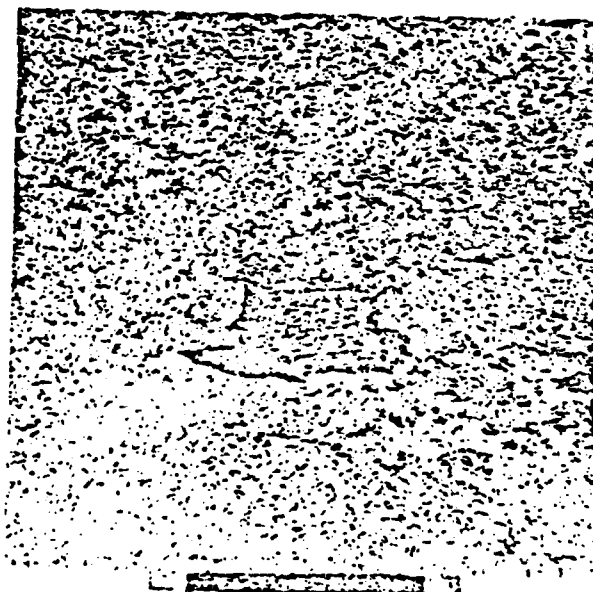


Figure 4. The noisy image with Poisson noise, with $\lambda = 0.05$.

and

$$\sigma_{R_D}^2 = E\left[\left(\frac{1}{\lambda}P[\lambda S] - S - u_{R_D}\right)^2\right] . \quad (14)$$

As before, if we treat S as an unknown constant equal to s , we obtain

$$\sigma_{R_D}^2|_{S=s} = s/\lambda \quad (15)$$

Thus an estimate of the signal, \hat{s} , may be obtained by using the relationship

$$\hat{s} = \lambda \hat{\sigma}_{R_D}^2 , \quad (16)$$

where $\hat{\sigma}_{R_D}^2$ is an estimate of the local variance of the difference image obtained using R_D for R in Eq.(9). The residual image obtained from the noisy observation of Fig. 4 is shown in Fig. 5. The estimate obtained from this difference image is shown in Fig. 6. This restored image demonstrates the ability of the estimator to extract the gross structure of the original image by making use of the signal information contained in the signal-dependent noise.

Signal Estimation from Multiple Observations

In some instances, such as television or movie imaging, multiple looks of a given object or scene are available. In such cases we may estimate the object over an ensemble of sample functions, instead of estimating from a single observation. Such an estimate does not suffer from the spatial smoothing effects that are present in many of the image estimation techniques. As a result one may expect a higher spatial resolution in restored images. Further, when the noise corrupting the signal contains a signal-dependent part, it is possible to obtain a second estimate by recovering the signal from the signal-dependent noise term. These issues are discussed in this section.

Consider an ensemble of noisy observations $\{R_i; i=1,2,\dots,M\}$ obtained when a signal S is degraded by a set of statistically independent noise terms N_{1i} and N_{2i} , with

$$R_i = S + f(S)N_{1i} + N_{2i}, \quad i=1,2,\dots,M. \quad (17)$$

If the noise processes are modeled as zero mean processes, then an estimate of the signal, \hat{S} , may be obtained by finding the sample mean over the ensemble at each pixel, using the equation

$$\hat{S} = \frac{1}{M} \sum_{i=1}^M R_i . \quad (18)$$

When such an estimate is subtracted from each of the noisy observations, the residual images, R_{Di} , consist mostly of noise terms as given by

$$R_{Di} = f(S)N_{1i} + N_{2i} . \quad (19)$$

Using the procedure described earlier, an estimate, \hat{S}_N , may be obtained from these residual images. Thus, if $\hat{\sigma}_{R_D}^2$ represents an estimate of the variance of the residual images at a given pixel, and if σ_1^2 and σ_2^2 are the variances of the noise terms N_1 and N_2 respectively, then an estimate \hat{S}_N is given by

$$\hat{S}_N = f^{-1} \left[\sqrt{\frac{\hat{\sigma}_{R_D}^2 - \sigma_2^2}{\sigma_1^2}} \right] , \quad (20)$$

where

$$\hat{\sigma}_{R_D}^2 = \frac{1}{M} \sum_{i=1}^M R_{Di}^2 - \left[\frac{1}{M} \sum_{j=1}^M R_{Dj} \right]^2 . \quad (21)$$

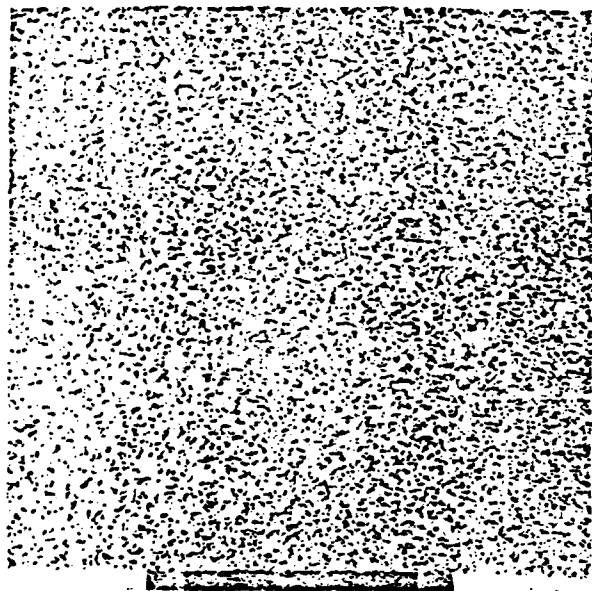


Figure 5. The difference image obtained from Fig. 4.

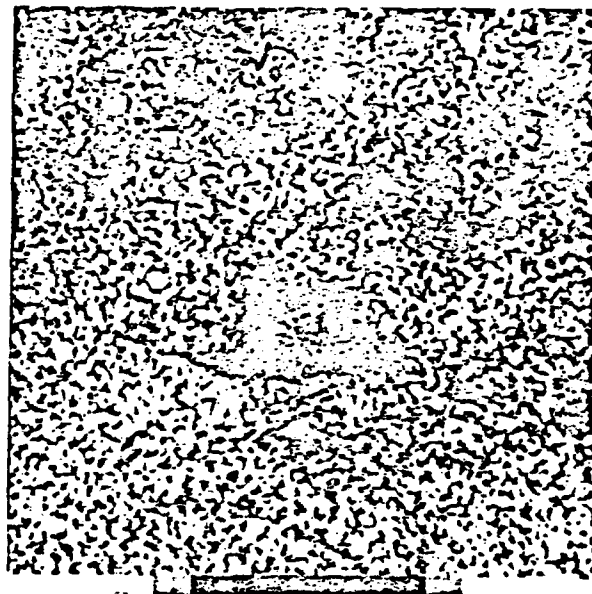


Figure 6. The image recovered from Fig. 4.

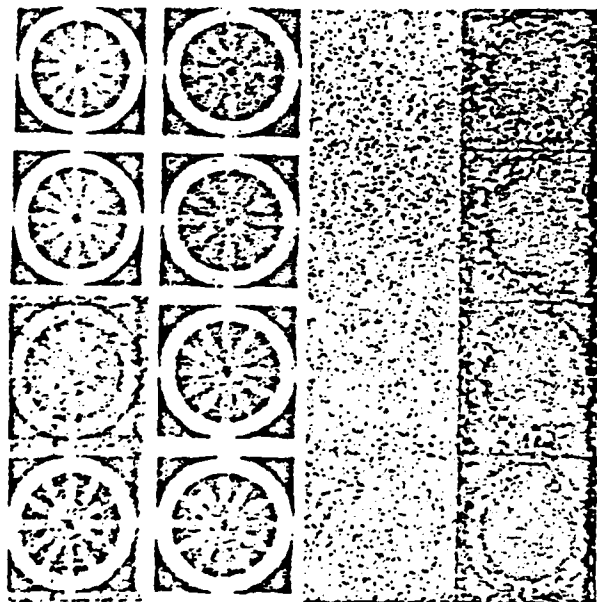


Figure 7. Image restoration using multiple observations.

a	e	i	m
b	f	j	n
c	g	k	o
d	h	l	p

Lay-out of the figure.

Note that the variables in Eqs. (17)-(21) represent the corresponding quantities at each pixel location and the same estimation procedure must be applied to each pixel to restore the entire image. To evaluate this estimation process, independent Gaussian noise samples were added to the original image of Fig. 7a, resulting in the four noisy images shown in Figs. 7e-h. The restored image obtained by finding the sample mean at each pixel is shown in Fig. 7b. The residual images obtained by subtracting the image of Fig. 7b from each of the noisy observations are shown in Figs. 7i-l (as before a constant background has been added to these images). The restored image obtained when Eq. (20) is applied to these images is shown in Fig. 7c. Note the remarkable detail visible in this restored image even when the sample size is as low as four. The signal recovered from each of the difference images shown in Figs. 7i,j,k using the spatial signal recovery techniques described earlier are shown in Figs. 7m,n,o. The estimate shown in Fig. 7p is the pixel-by-pixel mean of the estimates shown in Figs. 7m,n,o. The signal information may also be recovered using higher order moments. As an example, the restored image obtained by estimating the fourth moment of the difference images and other appropriate equations is shown in Fig. 7d. As expected, the use of multiple frame observations has completely eliminated the spatial smoothing problem encountered in single frame estimators. Further, these simulations have indicated that excellent noise suppression and the recovery of good quality images is possible even when the sample size is small.

Conclusions

The potential advantages of the procedures described for recovering a signal from signal-dependent noise have been discussed in this paper. It has been shown that it is possible to obtain a fairly good estimate of the signal even when all the signal information present in the observation is in the form of signal-dependent noise. Modifications to the estimators in order to recover a signal from multiple frame observations have also been presented. Although the techniques described are applicable to only Gaussian signal-dependent noise and Poisson noise, it is possible to extend the same principles to other types of signal-dependent noise, such as magnetic tape noise and reverberation noise.

Acknowledgments

This research was supported by the Joint Services Electronics Program at Texas Tech University.

References

1. P. G. Roetling, "Effects of Signal-dependent Granularity," J. Opt. Soc. Amer., Vol. 55, no. 1, pp. 67-71, 1965.
2. G. K. Froehlich, J. F. Walkup, and T. F. Krile, "Estimation in Signal-dependent Film-Grain Noise," Appl. Opt., Vol. 20, pp. 3619-3626, 1981.
3. R. Kasturi, Adaptive Image Restoration in Signal-Dependent Noise, Ph.D. Dissertation, Dept. of Elec. Eng., Texas Tech University, Lubbock, TX, 1982.
4. C. M. Lo, Estimation of Image Signals with Poisson Noise, Image Processing Institute, University of Southern California, Report no. 890, 1979.

POINTING AND TRACKING

THE PAGE

274

THE GEOMETRY OF SECOND ORDER LINEAR
PARTIAL DIFFERENTIAL OPERATORS

Gregory A. Fredricks

Department of Mathematics
Texas Tech University
Lubbock, Texas 79409

Texts on classical partial differential equations (e.g., Garabedian [6,p.57] and Courant and Hilbert [2,p.155]) always develop the well-known theory of canonical forms for second order linear partial differential equations in two variables. This micro-local result gives necessary and sufficient conditions for the reduction of the top order part of the equation to constant coefficients. The purpose of this paper is to investigate two macro-local problems on the reduction of the top order part of a linear partial differential operator to constant coefficients.

Section 1 contains the preliminaries and leads to statements of the two problems on the geometry of linear partial differential operators. The general results concerning second order linear partial differential operators are given in Section 2. Section 3 is devoted to some consequences of the theorem of Cotton, while Section 4 contains some miscellaneous remarks.

Section 1: Preliminaries

Let M be an n -dimensional smooth manifold and let $F(M, p; \mathbb{R}^m)$ denote the set of smooth maps from M to \mathbb{R}^m which are defined in a neighborhood of $p \in M$. If (x, U) is a chart of M , $p \in U$, and $\alpha = (\alpha_1, \dots, \alpha_n)$ is an n -tuple of nonnegative integers, we define $\partial f / \partial x^\alpha(p)$ for each $f \in F(M, p; \mathbb{R})$ by

$$(1.1) \quad \partial f / \partial x^\alpha(p) = D^\alpha(f \circ x^{-1})(x(p)),$$

where D^α denotes the usual partial derivative in \mathbb{R}^n . If $\alpha = (0, \dots, 0)$, then $D^\alpha g = g$. Define $|\alpha| = \alpha_1 + \dots + \alpha_n$ and $\alpha! = \alpha_1! \dots \alpha_n!$.

For each nonnegative integer r , we say that $f, g \in F(M, p; \mathbb{R}^m)$ are r -jet equivalent if for some chart (x, U) of M for which $p \in U$ we have

$$(1.2) \quad \partial f_i / \partial x^\alpha(p) = \partial g_i / \partial x^\alpha(p) \text{ for } i=1, \dots, m \text{ and } |\alpha| \leq r.$$

It follows from the chain rule that (1.2) is independent of the chart. The r -jet equivalence class of $f \in F(M, p; \mathbb{R}^m)$ is denoted by $j^r f(p)$ and the set of all r -jet equivalence classes in $F(M, p; \mathbb{R}^m)$ is denoted by $J^r(M, p; \mathbb{R}^m)$.

The set $J^r(M, p; \mathbb{R}^m)$ has a natural vector space structure inherited from $F(M, p; \mathbb{R}^m)$ and $J^r(M; \mathbb{R}^m) = \bigcup_{p \in M} J^r(M, p; \mathbb{R}^m)$ has a natural manifold structure which makes $J^r(M; \mathbb{R}^m)$ with its natural projection onto M into a vector bundle. We also consider jet bundles with a fixed target by defining

$$J^r(M, p; \mathbb{R}^m, 0) = \{j^r f(p) \mid f \in F(M, p; \mathbb{R}^m) \text{ and } f(p) = 0\}.$$

LINEAR PARTIAL DIFFERENTIAL OPERATORS

Proceeding as above, we obtain the vector bundle $J^r(M; R^m, 0)$ over M .

The case $m=1$ is of special interest. The dual bundle of $J^r(M; R, 0)$ is the r -th order tangent bundle $T^r(M)$ of Pohl [7]. Sections of $T^r(M)$ are r -th order partial differential operators on M (without constant term). In fact, charts of M induce local bases of sections for $J^r(M; R, 0)$ and $T^r(M)$ in a straightforward way. If (x, U) is a chart of M and if $x^\alpha = x_1^{\alpha_1} \dots x_n^{\alpha_n}$, then the maps $j^r x^\alpha$ for $0 < |\alpha| \leq r$, defined by

$$j^r x^\alpha(p) = j^r((x-x(p))^\alpha)(p) \text{ for } p \in U,$$

form the basis of sections of $J^r(M; R, 0)$ over U . The dual basis of sections of $T^r(M)$ over U is

$$\{(1/\alpha!) \partial / \partial x^\alpha \mid 0 < |\alpha| \leq r\}.$$

In a similar fashion, the dual bundle of $J^r(M; R)$ is $T^r(M) \oplus \mathbb{1}$, where $\mathbb{1}$ denotes the one-dimensional trivial bundle over M . Sections of $T^r(M) \oplus \mathbb{1}$ are r -th order partial differential operators on M with constant term.

We will now show that charts of M are equivalent to the existence of integrable, local sections of a certain fibre bundle. Let $RF(M, p; R^n)$ denote the subset of $f \in F(M, p; R^n)$ which are regular at p , i.e. for which the $n \times n$ Jacobian matrix of f relative to some (and hence any) chart of M near p is nonsingular at p . Let

$$RJ^r(M, p; R^n, 0) = \{j^r f(p) \mid f \in RF(M, p; R^n) \text{ and } f(p) = 0\}.$$

The set $RJ^r(M; R^n, 0) = \bigcup_{p \in M} RJ^r(M, p; R^n, 0)$ has a natural manifold

structure which makes $RJ^F(M; R^n, 0)$ with its natural projection onto M into a fibre bundle. A section s of $RJ^F(M; R^n, 0)$ is integrable if there exists a smooth map $f: M \rightarrow R^n$ which is regular on M such that

$$s(p) = j^F(f-f(p))(p) \text{ for each } p \in M.$$

We will denote such a section simply by $j^F f$.

Note that if (x, U) is a chart of M , then $j^F x$ is an integrable section of $RJ^F(M; R^n, 0)$ over U . In fact, if one identifies charts (x, U) and (y, V) for which $U=V$ and $x-y$ is a translation up to order r , then there is a natural one-to-one correspondence between charts of M and integrable, local sections of $RJ^F(M; R^n, 0)$.

We have already seen how a chart (x, U) of M , i.e. an integrable section $j^F x$ of $RJ^F(M; R^n, 0)$ over U , induces a local basis $\{\partial/\partial x^\alpha \mid 0 < |\alpha| \leq r\}$ of sections of $T^F(M)$ over U . These local bases are called coordinate frames for $T^F(M)$ over U .

Suppose now that s is a section of $RJ^F(M; R^n, 0)$ over U . For each $p \in U$ and each α with $0 < |\alpha| \leq r$, define

$$(1.3) \quad X_\alpha(p) = \partial/\partial x^\alpha(p) \text{ if } s(p) = j^F x(p).$$

The maps $\{X_\alpha \mid 0 < |\alpha| \leq r\}$ form a local basis of sections of $T^F(M)$ over U . Local bases which are induced from sections of $T^F(M)$ over U in this manner are called generalized coordinate frames for $T^F(M)$ over U . Note that generalized coordinate frames are induced by smoothly varying germs of coordinates.

Coordinate and generalized coordinate frames for $T^F(M) \otimes \mathbb{1}$, $J^F(M; R, 0)$ and $J^F(M; R)$ can be defined in the obvious ways. Because

LINEAR PARTIAL DIFFERENTIAL OPERATORS

of our interest in partial differential operators, we will only consider these frames for $T^r(M)$.

Since the geometry of partial differential operators is in the top order symbol, we can now state the two main problems on "reduction to constant coefficients" in the top order part. Let P be a section of $T^r(M)$ and let U be an open subset of M .

Problem 1: Find necessary and sufficient conditions on P and U for the existence of a coordinate frame $\{\partial/\partial x^\alpha\}$ for $T^r(M)$ over U and $b_\alpha \in \mathbb{R}$ for which

$$P = \sum_{|\alpha|=r} b_\alpha \partial/\partial x^\alpha + \dots \text{ on } U.$$

The symbolism "+..." means plus lower order terms.

Problem 2: Find necessary and sufficient conditions on P and U for the existence of a generalized coordinate frame $\{X_\alpha\}$ for $T^r(M)$ over U and $b_\alpha \in \mathbb{R}$ for which

$$P = \sum_{|\alpha|=r} b_\alpha X_\alpha + \dots \text{ on } U.$$

Note that the order of X_α is defined in the classical sense. Thus we see that the order of X_α is $|\alpha|$, but that X_α may also include lower order terms.

Since these problems are macro-local in nature, we will consider M to be an open subset U of \mathbb{R}^n . The following consequence of the chain rule is useful in translating these problems to problems in linear algebra and differential geometry. If $\{X_\alpha\}$ is a generalized coordinate frame for $T^r(U)$ induced by a section s of $RJ^r(U; \mathbb{R}^n, 0)$, then for each $p \in M$ and $0 < |\alpha| \leq r$ we have

$$D^2(p) = \sum_{0 < |\beta| \leq |\alpha|} \frac{1}{\beta!} D^2(x^\beta)(p) X_\beta(p),$$

where $s(p) = j^T x(p)$. Note that equality also holds in the preceding line if one sums over the indices $0 < |\beta| \leq r$. Thus, if P is a section of $T^r(U)$, say

$$P = \sum_{0 < |\alpha| \leq r} a_\alpha D^\alpha \quad \text{on } U,$$

with $a_\alpha: U \rightarrow \mathbb{R}$ smooth, then

$$(1.4) \quad P = \sum_{0 < |\beta| \leq r} \left(\sum_{0 < |\alpha| \leq r} a_\alpha g_{\alpha\beta} \right) X_\beta \quad \text{on } U,$$

where $g_{\alpha\beta}: U \rightarrow \mathbb{R}$ is defined by

$$g_{\alpha\beta}(p) = \frac{1}{\beta!} D^\alpha(x^\beta)(p) \quad \text{if } s(p) = j^T x(p).$$

Section 2: General results

It is convenient to use a different type of notation while considering Problems 1 and 2 for second order linear partial differential operators on an open subset U of \mathbb{R}^n . Let e_1, \dots, e_n denote the standard unit vectors in \mathbb{R}^n and define

$$D_i = D^{e_i} \quad \text{and} \quad D_{ij} = D^{e_i + e_j} \quad \text{for } i, j = 1, \dots, n.$$

In a similar fashion, we define $\partial/\partial x_i$ and $\partial/\partial x_{ij}$ for a coordinate frame $\{\partial/\partial x^\alpha\}$ for $T^2(U)$, and X_i and X_{ij} for a generalized coordinate frame $\{X_\alpha\}$ for $T^2(U)$.

Suppose now that P is a second order linear partial differential operator on U . Then P can be written uniquely in the form

$$(2.1) \quad P = \sum_{i,j} a_{ij} D_{ij} + \sum_i a_i D_i \quad \text{on } U,$$

LINEAR PARTIAL DIFFERENTIAL OPERATORS

where $a_{ij}, a_i: U \rightarrow \mathbb{R}$ are smooth functions and $a_{ij} = a_{ji}$. It follows from (1.4) that if $\{\partial/\partial x^a\}$ is a coordinate frame for $T^2(U)$ and

$$P = \sum_{k,l} b_{kl} \frac{\partial}{\partial x_{kl}} + \sum_k b_k \frac{\partial}{\partial x_k} \text{ on } U,$$

with $b_{kl} = b_{lk}$, then

$$b_{kl} = \sum_{i,j} a_{ij} D_i x_k D_j x_l \text{ on } U \text{ for } k, l = 1, \dots, n.$$

Thus the second order coefficients of P transform tensorially according to the matrix equation

$$Dx A^t Dx = B \text{ on } U,$$

where $A_{(i,j)} = a_{ij}$, $B_{(k,l)} = b_{kl}$ and Dx is the Jacobian matrix of the coordinates x . Given the symmetric matrix A with smooth entries, this equation is solvable for coordinates x on U and for a constant matrix B if and only if it is solvable for coordinates x on U and a constant matrix of the form

$$B_{p,q} = \text{diag} (1, \dots, 1, -1, \dots, -1, 0, \dots, 0),$$

with p 1's and q -1's, where p and q are nonnegative integers with $p+q \leq n$.

Thus we have

Problem 1 (r=2): Find necessary and sufficient conditions on P in (2.1) and U for the existence of coordinates x on U and nonnegative integers p and q for which

$$(2.2) \quad Dx A^t Dx = B_{p,q} \text{ on } U.$$

If $\{X_\alpha\}$ is a generalized coordinate frame for $T^2(U)$, then we obtain the relationship (2.2) where the coordinates x are germs of coordinates which vary smoothly on U . Thus we have

Problem 2 ($r=2$): Find necessary and sufficient conditions on P in (2.1) and U for the existence of a smooth map $G: U \rightarrow GL(n)$ and non-negative integers p and q for which

$$(2.3) \quad G A^T G = B_{p,q} \text{ on } U.$$

Note, from Sylvester's Theorem, that a necessary condition for solving (2.3) is that A has p positive and q negative eigenvalues at each point of U . Under this assumption we have (see [5] for more details) that

$$E = \{(u, g) \in U \times GL(n) \mid g A(u)^T g = B_{p,q}\}$$

is a submanifold of $U \times GL(n)$ and that (E, π, U) , with π denoting the obvious projection of E onto U , is a fibre bundle. Since the desired map G is a section of this fibre bundle, we can guarantee the existence of such a map by assuming, for example, that U is smoothly contractible (see Steenrod [9, p.53]). Hence we have

Theorem 2.4: Let U be a smoothly contractible open subset of \mathbb{R}^n and let P be given by (2.1). There exists a generalized coordinate frame $\{X_\alpha\}$ for $T^2(U)$ for which

$$P = \sum_{i=1}^p X_{ii} - \sum_{j=p+1}^{p+q} X_{jj} + \dots \text{ on } U$$

LINEAR PARTIAL DIFFERENTIAL OPERATORS

if and only if A has p positive and q negative eigenvalues at each point of U .

To solve (2.2) we need a smooth map $G: U \rightarrow GL(n)$ which satisfies (2.3) and for which there exists a solution to the system

$$D_j x_i = g_{ij} \text{ on } U \text{ for } i, j = 1, \dots, n,$$

where $G_{(i,j)} = g_{ij}$. From basic existence and uniqueness theorems for systems of first order partial differential equations we see that local solutions exist if and only if

$$D_k g_{ij} = D_j g_{ik} \text{ on } U \text{ for } i, j, k = 1, \dots, n.$$

If the matrix A is nonsingular on U , i.e., P is nondegenerate on U , then the preceding "integrability conditions" can be written as curvature conditions on A . (See Spivak [8, ch.4D] for more details.) Recall that the components R_{ijkl} of the Riemann curvature tensor corresponding to A are defined for $i, j, k, l = 1, \dots, n$ by

$$R_{ijkl} = (D_j a_{il} + D_i a_{jk} - D_l a_{ik} - D_k a_{jl})/2 \\ + \sum_{s,t} a_{st} ([il, s][jk, t] - [ik, s][jl, t]),$$

where $A_{(i,j)}^{-1} = a_{ij}$ and $[ij, k] = (D_j a_{ik} + D_i a_{jk} - D_k a_{ij})/2$.

As in the macro-local theory of canonical forms for second order linear partial differential equations (see Finn [4]), the "integrability conditions" are sufficient only on certain domains. The Poincaré lemma with compact supports can now be used to obtain the following macro-local version of a theorem due to Cotton [1].

AD-A137 619

ANNUAL REVIEW OF RESEARCH UNDER THE JOINT SERVICES
ELECTRONICS PROGRAM VO. (U) TEXAS TECH UNIV LUBBOCK
INST FOR ELECTRONIC SCIENCE R SAEKS ET AL. DEC 82
N00014-76-C-1136 F/G 9/3

4/4

UNCLASSIFIED

NL

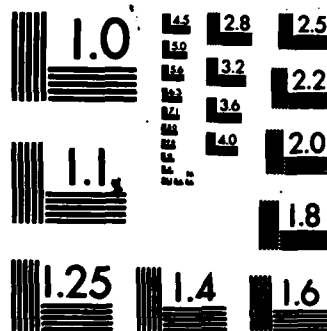
END

FILED

3

DEC

1982



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Theorem 2.5: Let \bar{U} be a smoothly contractible open subset of \mathbb{R}^n and let P , given by (2.1) on \bar{U} , be nondegenerate on \bar{U} . For each open subset U of \mathbb{R}^n with \bar{U} compact and contained in U there exists a coordinate frame $\{\partial/\partial x^a\}$ for $T^2(U)$ for which

$$P = \sum_{i=1}^p \partial/\partial x_{ii} - \sum_{j=p+1}^n \partial/\partial x_{jj} + \dots \text{ on } U$$

if and only if A has p positive eigenvalues at any point (and hence all points) of \bar{U} and the components R_{ijkl} of the Riemann curvature tensor corresponding to A vanish on \bar{U} .

Section 3: Consequences of the theorem of Cotton

We will assume throughout this section that \bar{U} and U are open subsets of \mathbb{R}^n with the properties stated in Theorem 2.5. Note (see Eisenhart [3, p.21]) that, due to various relationships between the components R_{ijkl} , there are $n^2(n^2-1)/12$ conditions in the vanishing of the Riemann curvature tensor corresponding to A .

We now begin a study of the case $n=2$, where we have a single vanishing condition on the curvature, namely, $R_{1221}=0$ on \bar{U} . Consider

$$(3.1) \quad P = aD_{11} + 2bD_{12} + cD_{22} + dD_1 + eD_2 \text{ on } \bar{U},$$

where $a, b, c, d, e: \bar{U} \rightarrow \mathbb{R}$ are smooth, and assume that P is nondegenerate on \bar{U} . We have

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \text{ and we let } A^{-1} = \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix}.$$

A calculation shows that R_{1221} vanishes on \bar{U} if and only if

LINEAR PARTIAL DIFFERENTIAL OPERATORS

$$(3.2) \quad \begin{aligned} &2\gamma_{11}-4\beta_{12}+2\alpha_{22}+a(2\alpha_1\beta_2-\alpha_1\gamma_1-\alpha_2^2)+c(2\beta_1\gamma_2-\alpha_2\gamma_2-\gamma_1^2) \\ &+b(\alpha_1\gamma_2-\alpha_2\gamma_1+4\beta_1\beta_2-2\beta_1\gamma_1-2\alpha_2\beta_2)=0 \text{ on } \bar{U}, \end{aligned}$$

where we have used f_i and f_{ij} to denote $D_i f$ and $D_{ij} f$, respectively.

Examination of the eigenvalue condition now gives the following consequence of Theorem 2.5.

Theorem 3.3: For P given in (3.1), there exist coordinates (x,y)
on U for which

$$(a) \quad P = \partial/\partial x^2 + \partial/\partial y^2 + \dots \text{ on } U;$$

$$(b) \quad P = -\partial/\partial x^2 - \partial/\partial y^2 + \dots \text{ on } U;$$

$$(c) \quad P = \partial/\partial x^2 - \partial/\partial y^2 + \dots \text{ on } U;$$

if and only if (3.2) holds and, respectively.

$$(a) \quad \det A > 0 \text{ on } \bar{U} \text{ and } a > 0 \text{ on } \bar{U};$$

$$(b) \quad \det A > 0 \text{ on } \bar{U} \text{ and } a < 0 \text{ on } \bar{U};$$

$$(c) \quad \det A < 0 \text{ on } \bar{U}.$$

We remark that, although condition (3.2) is rather long, it is generally much easier to check this condition in a specific example than it is to try to find coordinates which may or may not exist. This remains true even for more variables, where we may have a large number of conditions similar to (3.2).

Instead of using Theorem 3.3 to obtain results about specific operators P, we will assume that P has a special form and then apply Theorem 3.3 in that situation. For example, letting sgn denote the sign function, we have

Corollary 3.4: If $P=aD_{11}+cD_{22}+\dots$ on \bar{U} with a and c nonvanishing on \bar{U} , then there exist coordinates (x,y) on U for which

$$P = \operatorname{sgn}(a)\partial/\partial x^2 + \operatorname{sgn}(c)\partial/\partial y^2 + \dots \text{ on } U$$

if and only if

$$3a^3c_1^2+3c^3a_2^2-2a^3cc_{11}-2ac^3a_{22}-a^2ca_1c_1-ac^2a_2c_2=0 \text{ on } \bar{U}.$$

Corollary 3.5: If $P=aD_{11}+2bD_{12}+\dots$ on \bar{U} with b nonvanishing on \bar{U} , then there exist coordinates (x,y) on U for which

$$P = \partial/\partial x^2 - \partial/\partial y^2 + \dots \text{ on } U,$$

if and only if

$$b^2a_{11}-2abb_{11}-2b^2b_{12}+2bb_1b_2-3ba_1b_1+4ab_1^2=0 \text{ on } \bar{U}.$$

Instead of stating similar results for various forms of operators P , we will concentrate on situations where the general solution to the nonlinear partial differential equation (3.2) can be found. For example, letting $V^2=D_{11}+D_{22}$, we have

Corollary 3.6: If $P=aV^2+\dots$ on \bar{U} with a nonvanishing on \bar{U} , then there exist coordinates (x,y) on U for which

$$P = \operatorname{sgn}(a)(\partial/\partial x^2 + \partial/\partial y^2) + \dots \text{ on } U,$$

if and only if $V^2(\ln|a|) = 0$ on \bar{U} .

The proof entails letting $a=c$ in Corollary 3.4 and noting that

$$2a^3(aa_{11}-a_1^2+aa_{22}-a_2^2) = 2a^3V^2(\ln|a|).$$

LINEAR PARTIAL DIFFERENTIAL OPERATORS

Since every bounded harmonic function on \mathbb{R}^2 is constant, we obtain the following result.

Corollary 3.7: If $P = aV^2 + \dots$ on \mathbb{R}^2 with a nonvanishing and bounded on \mathbb{R}^2 , then there do not exist coordinates (x,y) on \mathbb{R}^2 in which the top order part of P has constant coefficients unless a itself is constant.

A result similar to Corollary 3.6 can be proven for $\square = D_{11} - D_{22}$ by noting that

$$2a^3(aa_{11} - a_1^2 - aa_{22} + a_2^2) = 2a^5 \square (\ln|a|)$$

and then using Corollary 3.4 to get

Corollary 3.8: If $P = a\square + \dots$ on \tilde{U} with a nonvanishing on \tilde{U} , then there exist coordinates (x,y) on U for which

$$P = \partial/\partial x^2 - \partial/\partial y^2 + \dots \text{ on } U,$$

if and only if $\square (\ln|a|) = 0$ on \tilde{U} , i.e., if and only if there exist smooth functions f and g for which

$$\ln|a(u,v)| = f(u+v) + g(u-v) \text{ for each } (u,v) \in \tilde{U}.$$

Here we use (u,v) to denote the canonical coordinates on \mathbb{R}^2 . A simple consequence of Theorem 3.3 (or Corollary 3.5) is the

Corollary: If $P = aD_{11} + 2D_{12} + \dots$ on \tilde{U} , then there exist coordinates (x,y) on U for which

$$P = \partial/\partial x^2 - \partial/\partial y^2 + \dots \text{ on } U,$$

if and only if there exist smooth functions f and g for which

$$a(u,v) = f(v)u + g(v) \text{ for each } (u,v) \in \bar{U}.$$

The next result follows from Theorem 3.3 after noting that

$$bb_{12} - b_1b_2 = (2a|b|)_{12}/b^2.$$

Corollary: If $P = 2bD_{12} + \dots$ on \bar{U} with b nonvanishing on \bar{U} , then there exist coordinates (x,y) on U for which

$$P = \partial/\partial x^2 - \partial/\partial y^2 + \dots \text{ on } U,$$

if and only if there exist smooth functions f and g for which

$$b(u,v) = f(u)g(v) \text{ for each } (u,v) \in \bar{U}.$$

We mention one final consequence of Theorem 3.3 because of its relationship to Laplace's operator in polar coordinates.

Corollary: If $P = D_{11} + cD_{22} + \dots$ with c nonvanishing on \bar{U} and depending only on u , then there exist coordinates (x,y) on U for which

$$P = \partial/\partial x^2 + \operatorname{sgn}(c) \partial/\partial y^2 + \dots \text{ on } U.$$

if and only if c , aside from a multiplicative constant, has the form

$$c(u) = (u-t)^{-2} \text{ for some } t \in \mathbb{R}.$$

We now consider some higher dimensional results. Since the number of conditions for the vanishing of the components R_{ijkl}

LINEAR PARTIAL DIFFERENTIAL OPERATORS

increases rapidly as n increases, results similar to those already given in this section become harder to obtain as n increases.

To get an idea of the conditions involved in the Riemann tensor we will state only one result in three dimensions.

Theorem: If $P = aD_{11} + bD_{22} + cD_{33} + \dots$ on \bar{U} with a, b, c nonvanishing on \bar{U} , then there exist coordinates (x, y, z) on U for which

$$P = \text{sgn}(a)\partial/\partial x^2 + \text{sgn}(b)\partial/\partial y^2 + \text{sgn}(c)\partial/\partial z^2 + \dots \text{ on } U,$$

if and only if the following hold on \bar{U} :

$$a^2 b a_1 b_1 + a b^2 a_2 b_2 - 3 b^3 a_2^2 + 2 a b^3 a_{22} - 3 a^3 b_1^2 + 2 a^3 b b_{11} - a b c a_3 b_3 = 0$$

$$a^2 c a_1 b_1 + a c^2 a_3 c_3 - 3 c^3 a_3^2 + 2 a c^3 a_{33} - 3 a^3 c_1^2 + 2 a^3 c c_{11} - a b c a_2 c_2 = 0$$

$$b^2 c b_2 c_2 + b c^2 b_3 c_3 - 3 c^3 b_3^2 + 2 b c^3 b_{33} - 3 b^3 c_2^2 + 2 b^3 c c_{22} - a b c b_1 c_1 = 0$$

$$2 a b c a_{23} + a b a_3 c_2 + a c a_2 b_3 - 3 b c a_2 a_3 = 0$$

$$2 a b c b_{13} + a b b_3 c_1 + b c a_3 b_1 - 3 a c b_1 b_3 = 0$$

$$2 a b c c_{12} + a c b_1 c_2 + b c a_2 c_1 - 3 a b c_1 c_2 = 0.$$

Since we will now turn to results which are true in n dimensions, we will temporarily cease using subscripts to denote partial derivatives.

Let u_1, \dots, u_n denote the canonical coordinates of R^n . Note that if $n \geq 2$ and $A = \text{diag}(a_1, \dots, a_n)$ with each a_m a function of u_m alone, then $[i, j] = 0$ for $i \neq j$ or $i \neq k$ and $[i, i] = (1/2a_i)'$.

We now see that $R_{ijkl} = 0$ for all i, j, k, l , and hence have the following consequence of Theorem 2.5.

Corollary: Let $P = a_1 D_{11} + \dots + a_n D_{nn} + \dots$ on \tilde{U} with a_1, \dots, a_n non-vanishing on \tilde{U} and each a_m a function of u_m alone. If p is the number of m for which $a_m > 0$ on \tilde{U} , then there exists a coordinate frame $(\partial/\partial x^a)$ for $T^2(U)$ for which

$$P = \sum_{i=1}^p \partial/\partial x_{1i} - \sum_{j=p+1}^n \partial/\partial x_{jj} + \dots \text{ on } U.$$

In Corollaries 3.6 and 3.8 we considered $P = aV^2 + \dots$ and $P = a\Box + \dots$, respectively. The corresponding cases in higher dimensions are covered in

Theorem 3.9: Suppose $n \geq 2$ and p is fixed with $0 < p < n$. Let $\delta_1, \dots, \delta_p$ equal one and let $\delta_{p+1}, \dots, \delta_n$ equal minus one. If $P = a(\delta_1 D_{11} + \dots + \delta_n D_{nn}) + \dots$ on \tilde{U} with a nonvanishing on \tilde{U} , then there exists a coordinate frame $(\partial/\partial x^a)$ for $T^2(U)$ for which

$$P = \text{sgn}(a)(\delta_1 \partial/\partial x_{11} + \dots + \delta_n \partial/\partial x_{nn}) + \dots \text{ on } U,$$

if and only if $|a|$ has one of the following forms on \tilde{U} :

$$r \left[\sum_{n=1}^n \delta_n (u_n - t_n)^2 \right]^2 \text{ or } \left[\sum_{n=1}^n q_n u_n + s \right]^2,$$

where $r, s \in \mathbb{R}$, $t = (t_1, \dots, t_n) \in \mathbb{R}^n$ and $q = (q_1, \dots, q_n) \in \mathbb{R}^n$ which satisfies

$$\sum_{n=1}^n \delta_n q_n^2 = 0.$$

Note that the conditions in the cases $p=n$ and $p=0$ are that a is constant or has the form $r|u-t|^4$ on \tilde{U} , where $r \in \mathbb{R}$ and $t \in \tilde{U}$.

LINEAR PARTIAL DIFFERENTIAL OPERATORS

Letting $v^2 = D_{11} + \dots + D_{nn}$, we now have the following higher dimensional analog of Corollary 3.7.

Corollary: If $P = av^2 + \dots$ on R^n with $n > 2$ and a nonvanishing on R^n , then there does not exist a coordinate frame $\{\partial/\partial x^a\}$ for $T^2(R^n)$ in which the top order part of P has constant coefficients unless a itself is constant.

Also note that if λ denotes the Lorentzian distance function, then the condition in the case $p=n-1$ of Theorem 3.9 is that a is of the form $[<q, u> + s]^2$, where $s \in R$ and q is in the light cone centered at the origin, or a is of the form $r[\lambda(u, t)]^4$, where $r \in R$ and the light cone centered at t does not meet \bar{U} .

We conclude this section with the

Proof of Theorem 3.9: Since $A = s \text{ diag } \{\delta_1, \dots, \delta_n\}$ and $A^{-1} = \text{diag } \{\delta_1, \dots, \delta_n\}/s$, it follows that $[i, j, k] = 0$ if i, j, k are distinct. One can now easily show that $R_{ijkl} = 0$ if i, j, k, l are distinct. Returning to the use of subscripts to denote partial differentiation, direct calculations show that

$$4a^3 R_{ijij} = \delta_j (2a_{ii} - a_i^2) + \delta_i (2a_{jj} - a_j^2) + \delta_i \delta_j \sum_{m=1}^n \delta_m a_m^2 \text{ if } i \neq j$$

and

$$4a^3 R_{ijil} = \delta_i (2a_{jl} - a_j a_l) \text{ if } i, j, l \text{ are distinct.}$$

Due to the various relationships between the components R_{ijkl} we see that the curvature conditions in Theorem 2.5 for A are

$$\begin{aligned}
 (3.10) \quad & \left\{ \begin{aligned} & \delta_i (2a_{ij} - a_i^2) + \delta_j (2a_{ji} - a_j^2) + r \delta_{ij} a_{ij}^2 = 0 \\ & \text{on } \bar{U} \text{ for } i < j. \end{aligned} \right. \\
 (3.11) \quad & 2a_{ij} - a_i a_j = 0
 \end{aligned}$$

We now assume that $a > 0$ on \bar{U} and note, in this case, that (3.11) is equivalent to $(a^{1/2})_{ij} = 0$ on \bar{U} . Since this holds for all $i < j$, it follows by integration that (3.11) holds for $a > 0$ if and only if there exist smooth functions f_n of u_n alone such that

$$a(u) = \left[\sum_{n=1}^N f_n(u_n) \right]^2 \text{ on } \bar{U}.$$

A calculation now shows that a function of this form satisfies

(3.10) if and only if

$$(3.12) \quad (r f_n) [\delta_i f_i'' + \delta_j f_j''] = r \delta_{ij} [f_n']^2 \text{ on } \bar{U} \text{ for } i < j.$$

We now see that $\delta_i f_i'' = \dots = \delta_n f_n''$ on \bar{U} and hence, since each f_n is a function of u_n alone, it follows that

$$(3.13) \quad \delta_i f_i'' = \dots = \delta_n f_n'' = 2r,$$

where $r \in \mathbb{R}$.

If $r = 0$, then f_n is a linear function of u_n and it follows from (3.12) that $a(u) = [c \langle q, u \rangle + s]^2$ on \bar{U} , where $s \in \mathbb{R}$ and $q \in \mathbb{R}^N$ which satisfies the condition given in the statement of the theorem.

Assume now that $r \neq 0$ and note from (3.13) that, for each $n = 1, \dots, N$, there exist $s_n, t_n \in \mathbb{R}$ such that

$$f_n(u_n) = r [\delta_n (u_n - t_n)^2 + s_n] \text{ on } \bar{U}.$$

It now follows from (3.12) that $\sum_{n=1}^N s_n = 0$ and hence we see that

LINEAR PARTIAL DIFFERENTIAL OPERATORS

$$a(u) = r^2 [\sum_m (\delta_m (u - c_m))^2]^2 \text{ on } \tilde{U}.$$

If $a < 0$ on \tilde{U} , then $|a| > 0$ on \tilde{U} and the preceding proof gives the desired result.

In concluding this section, we remark that since the top order part of P in (2.1) transforms according to a formula which is independent of the lower order coefficients, the results of Sections 2 and 3 also hold for quasi-linear operators.

Section 4: Miscellaneous remarks

The higher order cases of Problems 1 and 2 are much more difficult. In the case $r=3$ we see that if P is a third order linear partial differential operator on an open subset U of R^n , say,

$$P = \sum_{i,j,k} a_{ijk} D_{ijk} + \dots \text{ on } U,$$

with $a_{ijk}: U \rightarrow R$ smooth and symmetric in its indices, and if

$$P = \sum_{i,m,p} b_{imp} \partial/\partial x_{imp} + \dots \text{ on } U$$

for a coordinate frame $\{\partial/\partial x^a\}$ for $T^3(U)$ with b_{imp} symmetric in its indices, then

$$b_{imp} = \sum_{i,j,k} a_{ijk} D_i x_j D_k x_p \text{ on } U \text{ for } i,m,p=1,\dots,n.$$

As in the case $r=2$, we now see that if

$$A = \sum_{i,j,k} a_{ijk} \xi_i \xi_j \xi_k$$

is the symmetric cubic form associated to P , then Problem 1 [respectively, Problem 2] for $r=3$ can be stated as finding necessary and sufficient conditions for the existence of coordinates x on U [respectively, a smooth map $G: U \rightarrow GL(n)$] such that

$$D_x A = B \quad [\text{respectively, } G A = B] \quad \text{on } U,$$

where B is a real symmetric cubic form and the operation on the left is defined pointwise as the usual action of $GL(n)$ on the space of real symmetric cubic forms. We therefore have the obvious necessary condition on P that the real symmetric cubic forms $A(u)$ for each $u \in U$ belong to the same orbit of this action of $GL(n)$. Note that this condition is very strong, since it is well-known that there are an infinite number of such orbits. We suspect, however, that if U is smoothly contractible, then this constancy of orbit condition is necessary and sufficient for solving Problem 2. As in the case $r=2$, a solution to Problem 1 is provided by a solution G to Problem 2 which satisfies the usual integrability conditions. Transferring these integrability conditions to conditions on A in certain "generic" cases will involve the vanishing of some type of "higher order" curvature.

The preceding remarks carry over in the obvious way to the cases $r>3$. We remark, however, that the general problem of finding invariants which determine the orbits of the action of $GL(n)$ on the space of real symmetric r -linear forms is still open.

Finally, we remark that one can consider the problems of (total) reduction to constant coefficients by changing " $|\alpha|=r$ " to " $0 < |\alpha| \leq r$ " and deleting " $+\dots$ " in Problem 1 and 2. In [1], Cotton

LINEAR PARTIAL DIFFERENTIAL OPERATORS

considers this version of Problem 1 in the case $r=2$. To illustrate the form of these new problems, one can proceed in the case $r=2$ as in the beginning of Section 2 to obtain

Problem 1': Find necessary and sufficient conditions on P in (2.1) and U for the existence of coordinates x on U and $b_k, b_{kl} \in R$ for which

$$\begin{aligned} b_{kl} &= \sum_{i,j} a_{ij} D_i x_k D_j x_l \text{ on } U \text{ for } k, l=1, \dots, n, \text{ and} \\ (4.1) \quad b_k &= \sum_{i,j} a_{ij} D_{ij} x_k + \sum_i a_i D_i x_k \text{ on } U \text{ for } k=1, \dots, n. \end{aligned}$$

Problem 2': Find necessary and sufficient conditions on P in (2.1) and U for the existence of smooth maps $g_{ij}, g_{ijk}: U \rightarrow R$ with $g_{ijk} = g_{ikj}$ and $\det(g_{ij}) \neq 0$ on U and $b_k, b_{kl} \in R$ for which

$$\begin{aligned} b_{kl} &= \sum_{i,j} a_{ij} g_{ki} g_{lj} \text{ on } U \text{ for } k, l=1, \dots, n, \text{ and} \\ (4.2) \quad b_k &= \sum_{i,j} a_{ij} g_{kij} + \sum_i a_i g_{ki} \text{ on } U \text{ for } k=1, \dots, n. \end{aligned}$$

There are several ways to approach these problems. The approach to Problem 2' is to find a solution $G=(g_{ij}): U \rightarrow R$ to Problem 2 ($r=2$) and then solve the system of equations (4.2) for the unknown functions g_{kij} . Since this is now viewed as a system of n equations in $n \binom{n+1}{2}$ unknowns, we expect that solutions exist to Problem 2' under very mild conditions. The first approach to Problem 1' is to seek solutions to Problem 2' which satisfy the integrability conditions

$$D_k g_{ij} = D_j g_{ik} = g_{ijk} \text{ on } U \text{ for } i, j, k=1, \dots, n.$$

FREDRICKS

The second approach to Problem 1' is to find coordinates x on U which solve Problem 1 ($r=2$) and check to see if they satisfy (4.1). Note, however, that the b_k 's are uniquely determined by P and the coordinates x , and hence a study of Problem 1' under this approach would involve a study of systems of coordinates which preserve constant coefficients in the top order part.

In cases where $r > 2$, there is one exception to the preceding remarks. This is caused by the fact that, although the numbers of equations and unknowns in the system analogous to (4.2) increase as r increases, they do it in a sort of inverse way according to levels. Thus the subsystem corresponding to the part of order $r-1$ would be viewed as a system of $\binom{n+r-2}{r-1}$ equations in $n\binom{n+1}{2}$ unknowns. Thus we are led to the conclusion that there will be much stronger conditions for the existence of solutions to Problem 2'.

ACKNOWLEDGEMENTS

The author wishes to acknowledge many enlightening conversations with Aldo Andreotti and, at the same time, express deepest regrets at his untimely death on February 21, 1980.

This research was supported in part by the Office of Naval Research under Contract No. N00014-76-C-1136 conducted under the auspices of the Joint Services Electronics Program at Texas Tech University.

REFERENCES

- [1] É. Cotton, Sur les invariants différentiels de quelques équations linéaires aux dérivées partielles du second ordre, Ann. Sci. École Norm. Sup. 17 (1900), 211-244.

LINEAR PARTIAL DIFFERENTIAL OPERATORS

- [2] R. Courant and D. Hilbert, Methods of Mathematical Physics, Vol. II, Partial Differential Equations, Interscience, New York, 1962.
- [3] L.P. Eisenhart, Riemannian Geometry, Princeton Univ. Press, Princeton, 1926.
- [4] R. Finn, On the uniformization of plane direction fields, and of second-order partial differential operators, Proc. Camb. Phil. Soc. 73 (1973), 87-109.
- [5] G.A. Fredricks, Sylvester's theorem for matrices with smooth entries, to appear in Amer. Math. Monthly.
- [6] P.R. Garabedian, Partial Differential Equations, John Wiley and Sons, 1964.
- [7] W.F. Pohl, Differential geometry of higher order, Topology 1 (1962), 169-211.
- [8] M. Spivak, A Comprehensive Introduction to Differential Geometry, Vol. II, Publish or Perish, Boston, 1970.
- [9] N.E. Steenrod, The Topology of Fibre Bundles, Princeton Univ. Press, Princeton, 1951.

Received April 1982

Revised August 1982

END

FILMED

3-84

DTIC